2503012461

News & Numbers

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

News & Numbers

A GUIDE TO REPORTING STATISTICAL CLAIMS AND CONTROVERSIES IN HEALTH AND OTHER FIELDS

Victor Cohn

SENIOR WRITER AND COLUMNIST, FORMER SCIENCE EDITOR,
Washington Post

FOREWORD BY Frederick Mosteller

ROGER I. LEE PROFESSOR EMERITUS OF MATHEMATICAL STATISTICS,
Harvard University

A Project of the Center for Health Communication
Harvard School of Public Health

IOWA STATE UNIVERSITY PRESS / AMES

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

2503012462

RA407
.C64
1989

© 1989 Victor Cohn. All rights reserved

Composed by Iowa State University Press
Printed in the United States of America

No part of this book may be reproduced in any form or by any electronic or mechanical means, including information storage and retrieval systems, without written permission from the publisher, except for brief passages quoted in a review.

First edition, 1989

Library of Congress Cataloging-in-Publication Data

Cohn, Victor, 1919–
News & numbers.

"A project of the Center for Health Communication, Harvard School of Public Health."
1. Public health—Statistics. 2. Environmental health—Statistics. 3. Vital statistics. I. Harvard School of Public Health. Center for Health Communication. II. Title. III. Title: News and numbers.
RA407.C64 1989 362.1'021 88-6807
ISBN 0-8138-1442-1
ISBN 0-8138-1437-5 (pbk.)

A Note t

THE rule
fied. They a
stated or imp
porting, busi
This gu
language of
about th- ma
on soi
project of the
health and th
ciples and m
used by inqu
a scientific re
environment
weigh and
shows how t

25030124 63

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

A Note to Readers

THE rules of statistics are the rules of good thinking, codified. They apply to any kind of reporting in which numbers—stated or implied—are involved: political reporting, science reporting, business, economics, sports, or whatever.

This guide is an attempt to explain the role, logic, and language of statistics, so we reporters can ask better questions about the many alleged facts or findings that rest, or should rest, on some credible numbers. Because this manual began as a project of the Harvard School of Public Health, the reporting of health and the environment is the major example. But the principles and many of the suggested "questions for reporters" can be used by inquiring reporters in any field. They can help you read a scientific report or listen to the conflicting claims of politicians, environmentalists, physicians, scientists, or almost anyone and weigh and explain them. And the final chapter specifically shows how these principles apply in all areas.

VICTOR COHN

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

2503012464

Contents

FOREWORD BY *Frederick Mosteller*, ix

ACKNOWLEDGMENTS, xi

1. Facts and Figures—We Can Do Better, 3
2. The Certainty of Uncertainty, 8
3. The Scientific Way, 12
   - Probability, 14
   - "Power" and Numbers, 20
   - Bias and Confounders, 24
   - Variability, 30
4. Studies, Good and Bad, 35
   - Experiments versus Seductive Anecdotes, 37
   - Clinical Trials, 38
   - What Makes a Study Honest? 40
   - Epidemiology: Hippocrates to AIDS, 43
5. Questions Reporters Can Ask, 48
6. Tests and Testing, 64
   - Drugs and Drug Trials, 68
   - Animals as Models for Us, 72
7. Vital Statistics: The Numbers of Life and Health, 74
   - Crude Rates versus Rates That Compare, 76
   - Other Ways to Compare, 78
   - Reporting Hospital Death Rates, 79
   - Cancer Rates and Cancer "Cures", 86
   - The Important Questions about Cancer, 88
   - Shifts, Drifts, and Blips, 96

250301246S

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

8. The Statistics of Environment and Risk, 98
Who's Believable? 107
Questions to Ask, 108
Evaluating Environmental Hazards, 116
Advice from Reporters, 121

9. The Statistics of Politics, Economics, and Democracy, 126
The State of the Nation's Statistics, 146
The Bottom Line, 151

WHERE TO LEARN MORE: *A Bibliography and Other Sources,* 153

NOTES, 157

GLOSSARY/INDEX, 165

Foreword

REPORT
science to the
accuracy. Alth
stories, the bi
presents speci
misleading m
sistent," and "
sults that are
laymen
definiti
siderable diffe

Science
such as biosta
have been imp
ertheless, they
permanent fo

Victor C
ual to help all
wants to give
facts or mysti

Cohn's b
Science Policy
Research and
that faculty m
have been abl

25030 12466

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

Foreword

REPORTERS play an essential role in communicating science to the public. In common with scientists, they desire accuracy. Although health and medicine provide many exciting stories, the biostatistics that scientists must use in their studies presents special problems for reporters. It gives uncommon and misleading meanings to common words like "significant," "consistent," and "power." Mathematical statistics often produces results that are disturbingly counterintuitive, at least at first, to laymen and scientists alike. In vital statistics and epidemiology, definitions often seem arbitrary, and slight changes make considerable differences in the findings.

Science writers often take short courses in special topics such as biostatistics. I have taught in some of these courses and have been impressed by the seriousness of the participants. Nevertheless, they need some of this material in an accessible and permanent form.

Victor Cohn of the *Washington Post* has prepared this manual to help all reporters cut through these statistical tangles. He wants to give them a guide to the ways that statistics can clarify facts or mystify the reader.

Cohn's book grew out of the Media Project of our Health Science Policy Working Group of the Division of Health Policy Research and Education at Harvard University. I am pleased that faculty members of the Harvard School of Public Health have been able to help him produce this book as a visiting fellow

2503012467

Source: https://www.industrydocuments.ucsf.edu/docs/[illegible]

in 1978 and 1984 and as a contributor to the Health Science Policy Working Group.

Through the Media Project, with the help of Jay Winsten, we have also examined sources of pressures on the science writer.[1] In the future we want to use what we have learned through many discussions with science writers to advise scientists on their role in the media.

By such efforts, including this book, and by many similar efforts in this and other fields, scientists and writers may gradually upgrade the whole communication system, scientific and journalistic. Thus we may clear the communication channel between science and the public.

FREDERICK MOSTELLER

Acknowl

MY mai
has been Dr.
tus of mathe
partments of
Harvard Sch
energy, and
for the fact
approach rat
statem

T.
tion, and by
ing, which p
journalists, i

I did m
School of Pu
Center for I
guide, and
Thomas A.
were Drs. (
Kaiser orgar
and Peter M
writings I le
Cass Peterso
Runkle, no

I also o

25030124 68

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

Acknowledgments

MY main mentor and guide in the preparation of this book has been Dr. Frederick Mosteller, Roger I. Lee professor emeritus of mathematical statistics and former chairman of the departments of Biostatistics and Health Policy and Management, Harvard School of Public Health. He gave so fully of his time, energy, and knowledge that he should be listed as coauthor but for the fact that I sometimes used a journalist's freewheeling approach rather than a statistician's rigor. This makes any misstatements mine.

The project was supported by the Russell Sage Foundation, and by the Council for the Advancement of Science Writing, which pointed the way in holding seminars on statistics for journalists, including the first of its kind in 1964.

I did much of the work as a visiting fellow at the Harvard School of Public Health, where Dr. Jay Winsten, director of the Center for Health Communication, was another indispensable guide, and Drs. John Bailar III, Nan Laird, Philip Lavin, Thomas A. Louis, and Marvin Zelen were valuable helpers. As were Drs. Gary D. Friedman and Thomas M. Vogt of the Kaiser organizations, Michael Greenberg of Rutgers University, and Peter Montague of Princeton University (on all of whose writings I leaned); Lewis Cope of the Minneapolis *Star Tribune;* Cass Peterson of the *Washington Post;* and my daughter, Deborah Runkle, no mean statistician.

I also owe thanks to Harvard's Drs. Peter Braun, Harvey

2503012469

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

Fineberg, Howard Frazier, Howard Hiatt, William Hsaio, Herb Sherman, and William Stason. And to Drs. Stuart A. Bessler, Syntex Corporation; H. Jack Geiger, City University of New York; Nicole Schupf Geiger, Manhattanville College; Charles Moertel, Mayo Clinic; Arnold Relman, *New England Journal of Medicine;* Eugene Robin, Stanford University; and Sidney Wolfe, Public Citizen Health Research Group. Also Katherine Wallman, Council of Professional Associations on Federal Statistics; Howard L. Lewis, American Heart Association; Philip Meyer, University of North Carolina; Mildred Spencer Sanes; Earl Ubell, WCBS-TV, New York City; and Philip Hilts, Cristine Russell, and Barry Sussman, *Washington Post.* I am indebted to my editors at the *Washington Post,* particularly Abigail Trafford, Ben Cason, Carol Krucoff, Len Downie, and Howard Simons for their understanding and support.

The work was also aided by the Andrew W. Mellon Foundation. The American Cancer Society, American Heart Association, Commonwealth Fund, Gannett Foundation, Henry J. Kaiser Family Foundation, Mayo Medical Resources, Milbank Memorial Fund, Pew Charitable Trusts, Philip L. Graham Fund, Russell Sage Foundation, and John Cowles, Jr., have contributed to this manual's initial distribution.

Ne
Ni

2503012470

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

News & Numbers

2503012471

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

Facts and Figures— We Can Do Better 1

Facts and Figures! Put 'em Down!

—Charles Dickens (in *The Chimes*)

There are lies, there are damned lies, and there are statistics.

—Disraeli

Almost everyone has heard that "figures don't lie, but liars can figure." We need statistics, but liars give them a bad name, so to be able to tell the liars from the statisticians is crucial.

—Dr. Robert Hooke

WE journalists like to think we deal mainly in facts and ideas, but much of what we report is based on numbers.

Politics comes down to votes. Budgets and dollars dominate government. The economy, business, employment, sports—all demand numbers.

The environment, pollutants, toxic chemicals. Again, we see counts and measurements and, most likely, widely varying estimates, some careful, some questionably high or low. An environmentalist says a nuclear power plant or toxic waste dump will cause so many cases of cancer. An industry spokesman denies it. What are their numbers? Where did they get them? How valid are they?

A doctor reports a promising, even exciting new treatment. Is the claim justified or based on a biased or unrepresentative sample? Or too few patients to justify any claim? Science, medicine, technology, the weather, intelligence—all are statistical.

250301247 2

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

Science is observation, experimentation, measurement, and all these involve numbers, whether we reporters pay attention to them or not.

Statistics are used or misused even by people who tell us, "I don't believe in statistics," then claim that all of us or most people or many do such and such. The question for reporters is, how should we not merely repeat such numbers, stated or implied, but also interpret them to deliver the best possible picture of reality?

We can be better reporters if we understand how the best statisticians—the best figurers—figure. And if we learn a few questions to help us separate the wheat from the chaff.

I do not say that telling the truth—describing reality—will then become easy, for we are constantly bombarded with sweeping claims in convincing wrappings, and the disputed subjects are endless. Medical and surgical treatments, radiation, pesticides, nuclear power, the probability of environmental disasters, the side effects of medicines—almost nothing seems settled.

Like it or not, we must wade in. Whether we will it or not, we have in effect become part of the regulatory apparatus. Dr. Peter Montague of Princeton University tells us, "The environmental and toxic situation is so complex, we can't possibly have enough officials to monitor it. Reporters help officials decide where to focus their activity."

"Journalists opened up" the Love Canal toxic waste issue by "independent investigation," according to Cornell University's Dr. Dorothy Nelkin. "The extensive press coverage contributed to investigations that eventually forced the re-staffing of the Environmental Protection Agency and the creation of a national toxic waste disposal program."[1]

That very coverage, however, may also have stampeded public officials into hasty, ill-conceived studies that left unanswered the crucial question: Did the Love Canal wastes actually cause birth defects and other physical problems?[2] The very way we report a medical or environmental controversy can affect the outcome. If we ignore a bad situation, the public may

suffer. If we v
"no danger," th
experimental
false hope.

It is not
National Can
refuse to cor
think "carcin
persons prob
cancers are e
most informe
related main
and very pos
percent of al
carcinogens—
foods, air, an

When it
issues, or wh
making the s
state or unde
of *Pr*
statist.
terpretation.
evident; you
negative]. A
sterile is mor
that apple p

We also
the space or
news direct
story yet." E
done. In a
major south
traction afte
who worked
numbers fr

2503012473

Source: https://www.industrydocuments.ucsf.edu/docs/

suffer. If we write "danger," the public may quake. If we write "no danger," the public may be falsely reassured. If we paint an experimental medical treatment too brightly, the public is given false hope.

It is not just what we write, it is what we emphasize. A National Cancer Institute survey indicated that many persons refuse to consider healthy changes in life-style because they think "carcinogens are everywhere in the environment." Such persons probably have read or heard again and again that most cancers are environmentally related, although, in the opinion of most informed scientists, most fatal "environmental" cancers are related mainly to individual behavior, outstandingly smoking, and very possibly diet. By various estimates, perhaps 5 to 15 percent of all cancers are related to exposures to man-made carcinogens—chemicals we have inserted into the workplace, foods, air, and water.[3]

When it comes to such emotionally charged and complex issues, or when it simply comes to running for page one or making the six o'clock news, the best among us sometimes overstate or understate. Philip Meyer, veteran reporter and author of *Precision Journalism*, writes, "Journalists who misinterpret statistical data usually tend to err in the direction of overinterpretation. . . . The reason for this professional bias is self-evident; you usually can't write a snappy lead upholding [the negative]. A story purporting to show that apple pie makes you sterile is more interesting than one that says there is no evidence that apple pie changes your life."[4]

We also work fast, sometimes too fast, with severe limits on the space or time we may fill. We find it hard to tell editors or news directors, "I haven't had enough time. I don't have the story yet." Even a long-term project or special may be hurriedly done. In a newsroom "long-term" may mean a few weeks. A major southern newspaper had to print a long, front-page retraction after a series of front-page stories alleged that people who worked at or lived near a plutonium plant suffered in excess numbers from a blood disease. "Our reporters obviously had

2503012474

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

confused statistics and scientific data," the editor admitted. "We did not ask enough questions."[5]

We tend to oversimplify. We may report, "A study showed that black is white" or "So-and-so announced that . . . ," when a study merely suggested that there was some evidence that such might be the case. We may slight or omit the fact that a scientist calls a result "preliminary." As scientific unsophisticates, we may confuse a study that merely suggests a hypothesis that should be investigated—very frequently the case—with a study that presents strong and conclusive evidence.

We often omit essential perspective, context, or background. Dr. Thomas Vogt of the Kaiser Permanente Center for Health Research tells of seeing the headline "Heart Attacks From Lack of 'C'" and then, two months later, "People Who Take Vitamin C Increase Their Chances of a Heart Attack."[6] Both stories were based on limited, and far from conclusive, animal studies.

Scientists who do poor studies or overstate their results deserve part of the blame. But bad science is no excuse for bad journalism. We tend to rely most on "authorities" who are either most quotable or quickly available or both, and they often tend to be those who get most carried away with their sketchy and unconfirmed but "exciting" data—or have big axes to grind, however lofty their motives. The cautious, unbiased scientist who says, "Our results are inconclusive" or "We don't have enough data yet to make any strong statement" or "I don't know" tends to be omitted or buried someplace down in the story.

We are influenced too by intense and growing competition to tell the story first and tell it most dramatically. I was once asked by a Harvard researcher, "Does competition affect the way you present a story?" I thought and had to answer, "We have to almost overstate. We have to come as close as we can within the boundaries of truth to a dramatic, compelling statement. A weak statement will go no place." Another reporter said, "The fact is, you are going for the strong [lead and story]. And, while

2503012475

Source: https://www.industrydocuments.ucsf.edu/docs/[illegible]

not patently absurd, it may not be the lead you would go for a year later."[7]

We reporters are also subject to human hope and human fear. A new "cure" comes along, and we want to believe it. A new alarm is sounded, and we too tremble.

Alarms also make news. We too often obey a sardonic maxim: Bad news is good news; good news is no news. Dr. H. Jack Geiger, a respected former science writer and now a professor of medicine, says,

> I know I wrote stories in which I explained or interpreted the results wrongly. I wrote stories that didn't have the disclaimers I should have written. I wrote stories under competitive pressure, when it became clear later that I shouldn't have written them. I wrote stories when I hadn't asked—because I didn't know enough to ask—"Was your study capable of getting the answers you wanted? Could it be interpreted to say something else? Did you take into account possible confounding factors?"

How can we learn to do better?

2503012476

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

2 The Certainty of Uncertainty

Too much of the science reporting in the press [blurs] what we're sure of and what we're not very sure of and what is inconclusive. The notion of tentativeness tends to drop out of much reporting.

—Dr. Harvey Brooks

The only trouble with a sure thing is the uncertainty.

—Author unknown

THE first thing to understand about science is that it is almost always uncertain. A scientist, seeking to explain or understand something—be it the behavior of an atom or the effect of the toxic chemicals at a Love Canal—usually proposes a hypothesis, then seeks to test it by experiment or observation. If the evidence is strongly supportive, the hypothesis may then become a theory or at some point even a law, like the law of gravity.

A theory may be so solid that it is generally accepted. Example: the theory that cigarette smoking causes lung cancer, for which almost any reasonable person would say the case has been proved, for all practical purposes. The phrase "for all practical purposes" is important, for scientists, being practical people, must often speak at two levels: the strictly scientific level and the level of ordinary reason that we require for daily guidance.

Example: In June 1985, 16 forensic experts examined the bones that were supposedly those of the "Angel of Death," Dr. Josef Mengele. Dr. Lowell Levine, delegated by the Department of Justice, then said, "The skeleton is that of Josef

Mengele withi
cos Segre of t
with the law c
cians." Pushed
important mat
of the patholo
findings.[1] (Lat

But all ar
tainty in almo
bility that such

Widely b
wholly or part
say," reports I
Journal of Medi
help the publi
with an eleme
a probable na
not certainty. \
best opinion a
future."

Exa..
mended that v
cal cancer. Th
three years fo
Statistics had
matter is still
changed again

Scientists
a failing. Whe
sionally shows
right, the scie
ing.

The publ
have a hard t
sions. We all
today and anc
show discussi

2503012477

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

Mengele within a reasonable scientific certainty," and Dr. Marcos Segre of the University of Sao Paulo, explained, "We deal with the law of probabilities. We are scientists and not magicians." Pushed by reporters' questions—after all, this was an important matter, and what should the public believe?—several of the pathologists said they had "absolutely no doubt" of their findings.[1] (Later evidence made the case even stronger.)

But all any scientist can scientifically say—say with certainty in almost any such case—is, there is a very strong probability that such and such is true.

Widely believed theories or conclusions are often proved wholly or partly wrong. "When it comes to almost anything we say," reports Dr. Arnold Relman, editor of the *New England Journal of Medicine*, "you, the reporter, must realize—and must help the public understand—that we are almost always dealing with an element of uncertainty. Most scientific information is of a probable nature, and we are only talking about probabilities, not certainty. What we are concluding is the best we can do, our best opinion at the moment, and things may be updated in the future."

Example: Until 1980 the American Cancer Society recommended that women have an annual Pap smear to detect cervical cancer. The recommendation was then changed to every three years for many women, after two initial examinations. Statistics had shown that this would be equally effective.[2] The matter is still controversial, and the recommendation has been changed again in the light of new knowledge.

Scientists are often wrong. In science this is not necessarily a failing. When new evidence disproves an old theory, or occasionally shows that some little believed, even kooky notion is right, the scientific method is doing what it should. It is working.

The public, and even some reporters and especially editors, have a hard time understanding these sometimes drastic revisions. We all hear the question, Why do they say one thing today and another thing tomorrow? I was once on a radio talk show discussing unsettled medical controversies when a testy

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

2503012478

listener phoned in to exclaim, " 'They say' is a damned liar!"

"They" of course may be different theys who arrive at different conclusions about inconclusive evidence in a thousand areas: the role of fats and cholesterol in the diet, the effects of low-level radioactivity, the cause of the extinction of dinosaurs.

Why so much uncertainty? Science is always a continuing story. Nature is complex, and almost all methods of observation and experiment are imperfect. "There are flaws in all studies," says Harvard's Dr. Marvin Zelen.[3] There may be weaknesses, often unavoidable ones, in the way a study is designed or conducted. Observers are subject to human bias and error. Subjects fluctuate. Measurements fluctuate.

Many studies are thus inconclusive, and virtually no single study proves anything. "Fundamentally," writes Dr. Thomas Vogt, "all scientific investigations require confirmation, and until it is forthcoming all results, no matter how sound they may seem, are preliminary."[4]

Medicine, in particular, is full of disagreement and controversy. "No clinical trial is ever perfect," Harvard's Dr. John Bailar observes. Unlike new drugs, medical treatments and tests and surgical operations need not even be subjected to experimental studies before being applied. "Most treatments escape and will continue to escape rigorous evaluation," Bailar says.[5]

The reasons are many: lack of funds to mount enough trials; lack of enough patients at any one center to mount a meaningful trial; the expense and difficulty of doing multicenter trials; the swift evolution and obsolescence of medical techniques; the fact that, with the best of intentions, medical data—histories, physical examinations, interpretations of tests, descriptions of symptoms and diseases—are notoriously inexact and vary from physician to physician; and the serious ethical obstacles to trying a new procedure when an old one is doing some good, or to experimenting on children, pregnant women, or the mentally ill.

While all studies have flaws, some have more flaws than others. Study after study has found that many articles in the most prestigious medical journals are replete with shaky statis-

tics and lack o
tients' complic
up. Papers pre
reported by th
mere progress
tive results tha
or criticism o
uncertain find

The upsh
organization's
care is based
. . Seemingly
doctrines, per
out to be supp
be found."[7]

In gener
possible bene
that only a r
cancer. Only
less dra
omy, o
is rich in
or statistically
discarded.

Occasion
sults. More o
data that con
tical method
ascribing frau
in mind the
competence t

So some
tainty need
survive on t
policy, to go
basis of incor
can do so.

2503012479

Source: https://www.industrydocuments.ucsf.edu/docs/gnpl0000

tics and lack of any explanation of such crucial matters as patients' complications and the number of patients lost to follow-up. Papers presented at medical meetings, many of them widely reported by the media, are even less reliable. Many papers are mere progress reports on incomplete studies. Some state tentative results that later collapse. Some are given to draw comment or criticism or get others interested in a provocative but still uncertain finding.[6]

The upshot, according to Dr. Gary Friedman of the Kaiser organization's Permanente Medical Group: "Much of health care is based on tenuous evidence and incomplete knowledge. . . . Seemingly authoritative statements and accepted medical doctrines, perpetuated through textbook and lectures, often turn out to be supported by the most meager of evidence, if any can be found."[7]

In general, possible risks tend to be underestimated and possible benefits overestimated. For decades surgeons swore that only a radical mastectomy was the treatment for breast cancer. Only recently were clinical trials mounted to show that less drastic treatments seem equally effective. Prefrontal lobotomy, overstrict bed rest, drugs by the carload—medical history is rich in treatments that were given for years without question or statistically rigorous study, only to be proved wrong and discarded.

Occasionally, unscrupulous investigators falsify their results. More often, they may wittingly or unwittingly play down data that contradict their theories, or they may search out statistical methods that give them the results they want. Before ascribing fraud, says Harvard's Dr. Frederick Mosteller, "keep in mind the old saying that most institutions have enough incompetence to explain almost any results."[8]

So some uncertainty almost always prevails. But uncertainty need not stand in the way of good sense. To live—to survive on this globe, to maintain our health, to set public policy, to govern ourselves—we almost always must act on the basis of incomplete or uncertain information. There is a way we can do so.

2503012480

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

3 The Scientific Way

Somehow the wondrous promise of the earth is that there are things beautiful in it, things wondrous and alluring, and by virtue of your trade, you want to understand them.

—Mitchell Feigenbaum
Cornell University physicist and mathematician

The great tragedy of Science—the slaying of a beautiful hypothesis by an ugly fact.

—Thomas Henry Huxley

TO reporters, the world is full of true believers, peddling their "truths." The sincerely misguided and the outright fakers are often highly convincing, also newsy. How can we tell the facts, or the probable facts, from the chaff?

We can borrow from science. We can try to judge all possible claims of fact by the same methods and rules of evidence that scientists use to derive some reasonable guidance in scores of unsettled issues.

As a start, we can ask these questions:

How do you know?

Have the claims been subjected to any studies or experiments?

Were the studies acceptable ones, by general agreement? For example: Were they without any substantial bias?

Have results been fairly consistent from study to study?

Have the findings resulted in a consensus among others in the same field? Do at least the majority of informed persons agree? Or should we withhold judgment until there is more evidence?

Always: *Are the conclusions backed by believable statistical evidence?*

2503012481

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

And what is the degree of certainty or uncertainty? How sure can you be?

Obviously, much of statistics involves attitude or policy rather than numbers. And much, at least much of the statistics that reporters can most readily apply, is good sense.

There are many definitions of statistics as a tool. A few useful ones: The science and art of gathering, analyzing, and interpreting data; a means of deciding whether an effect is real; a way of extracting information from a mass of raw data; a set of mathematical processes derived from probability theory.

Statistics can be manipulated by charlatans, self-deluders, and inexpert statisticians. Deciding on the truth of a matter can be difficult for the best statisticians, and sometimes no decision is possible. Uncertainty will ever rule in some situations and lurk in almost all.

There are rare situations in which no statistics are needed. "Edison had it easy," says Dr. Robert Hooke, a statistician and author. "It doesn't take statistics to see that a light has come on."[1] It did not take statistics to tell 19th-century physicians that Morton's ether anesthesia permitted painless surgery or to tell 20th-century physicians that the first antibiotics cured infections that until then had been highly fatal.

Overwhelmingly, however, the use of statistics, based on probability, is called the soundest method of decision making, and the use of large numbers of cases, statistically analyzed, is called the only means for determining the unknown cause of many events. Birth control pills were tested on several hundred women, yet the pills had to be used for several years by millions before it became unequivocally clear that some women would develop heart attacks or strokes. The pills had to be used for some years more before it became clear that the greatest risk was to women who smoked and women over 35.

The best statisticians, let alone practitioners on the firing line (for example, physicians), often have trouble deciding when a study is adequate or meaningful. Most of us cannot become

25030124B2

Source: https://www.industrydocuments.ucsf.edu/docs/...

statisticians, but we can at least learn that there are studies and studies, and the unadorned claim "We made a study" or "We did an experiment" may not mean much. We can learn to ask more pointed questions if we understand some basic concepts and other facts about scientific studies.

These are some bedrock statistical concepts:

- Probability
- "Power" and numbers
- Bias and confounders
- Variability

Probability

Scientists cope with uncertainty by measuring probabilities. Since all experimental results and all events can be influenced by chance and almost nothing is 100 percent certain in science and medicine and life, probabilities sensibly describe what has happened and should happen in the future under similar conditions. Aristotle said, "The probable is what usually happens," but he might have added that the improbable happens more often than most of us realize.

The accepted numerical expression of probability in evaluating scientific and medical studies is the *P* (or *probability*) value. The *P* value is one of the most important figures a reporter should look for. It is determined by a statistical formula that takes into account the numbers of subjects or events being compared in order to answer the question, could a difference or result this great or greater have occurred *by chance alone?* By more precise definition, the *P* value expresses the probability that an observed relationship or effect or result could have *seemed* to occur by chance *if there had actually been no real effect.* A low *P* value means a low probability that this happened, that a medical treatment, for example, might have been declared beneficial when in truth it was not.

Here is why the *P* value is used to evaluate results. A

scientific inves
commonly sets
hypothesis: that
back the origi
pothesis. The
number or as
"greater than"
pened, that th
chance—or, to
variation.

• By con
only 5 or few
pened by cha
usually called
ues are used).
ally implies th

• *A highe*
statistically si
result is due

Ir
the sho
dinary logic.
replaces "it c

Why the
People have
purposes. Ar
Mosteller tel
class and aft
cious going
the fifth hea
chance in 16
that five hea
there is some
neighborhoo

Another
late a *confide*

25030124B3

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

scientific investigator first forms a hypothesis. Then he or she commonly sets out to try to *disprove* it by what is called the *null hypothesis:* that there is no effect, that nothing will happen. To back the original hypothesis, the results must *reject* the null hypothesis. The *P* value, then, is expressed either as an exact number or as $<.05$, say, or $>.05$, meaning "less than" or "greater than" a 5 percent probability that nothing has happened, that the observed result could have happened just by chance—or, to use a more elegant statistician's phrase, by *random variation.*

- By convention, *a P value of .05 or less,* meaning there are only 5 or fewer chances in 100 that the result could have happened by chance, is most often regarded as *low.* This value is usually called *statistically significant* (though sometimes other values are used). The unadorned term "statistically significant" usually implies that *P* is .05 or less.
- *A higher P value,* one *greater than .05,* is usually seen as not statistically significant. The higher the value, the more likely the result is due to chance.

In common language, a low chance of chance alone calling the shots replaces the "it's certain" or "close to certain" of ordinary logic. A strong chance that chance could have ruled replaces "it can't be" or "almost certainly can't be."

Why the number .05 or less? Partly for standardization. People have agreed that this is a good cutoff point for most purposes. And partly out of old friend common sense. Frederick Mosteller tells us that if you toss a coin repeatedly in a college class and after each toss ask the class if there is anything suspicious going on, "hands suddenly go up all over the room" after the fifth head or tail in a row. There happens to be only 1 chance in 16—.0625, not far from .05, or 5 chances in 100—that five heads or tails in a row will show up in five tosses, "so there is some empirical evidence that the rarity of events in the neighborhood of .05 begins to set people's teeth on edge."[2]

Another common way of reporting probability is to calculate a *confidence level,* as well as a *confidence interval* (or *confidence*

2503012484

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

limits or *range*). This is what happens when a political pollster reports that candidate X would now get 50 percent of the vote and thereby lead candidate Y by 3 percentage points, "with a 3-percentage-point margin of error plus or minus and a 95 percent confidence level." In other words, Mr. or Ms. Pollster is 95 percent confident that X's share of the vote would be someplace between 53 and 47 percent. Similarly, candidate Y's share might be 3 percentage points greater (or less) than the figure predicted. In a close election, that margin of error could obviously turn a predicted defeat into victory. And that sometimes happens.

An important point in looking at the results of political polls (and any other statements of confidence): In the reports we read, the plus or minus 3 (or whatever) percentage points is often omitted, and the pollster merely mentions a "3-point margin of error." This means there is actually a 6-point range within which the truth *probably* lurks.

The more people who are questioned in a political poll or the larger the number of subjects in a medical study, the greater the chance of a high confidence level and a narrow, and therefore more reassuring, confidence interval.

No matter how reassuring they sound, *P* values and confidence statements cannot be taken as gospel, for .05 is not a guarantee, just a number. There are several important reasons for this.

• All that *P* values measure is the *probability* that the results might have been produced by some sneaky random process. In 20 results where only chance is at work, 1, on the average, will have a reassuring-sounding but misleading *P* value of $< .05$. One, in short, may be a false positive.

Dr. Marvin Zelen points out that there may be 6,000 to 10,000 clinical (medical) trials of cancer treatment under way today, and if the conventional value of .05 is adopted as the upper permissible limit for false positives, then every 100 studies with no actual benefit may, on average, produce 5 false-positive results. Hence, we may expect 50 false positive results, on

average, for ever
fact has said, "W
chemotherapy ir
therapies in the
paths.

Amazingly,
tected. Scientist
negative results
them. Nor are s
ing studies that
firmatory studie

• Statistica
cause and effec
member the roc
Unless an assoc
that the case is
ing more study

To statisti
ference betwee
there is
correlation

• If the nu
value may sin
detect somethi
jects. Highly "s
ble differences

• An imp
other variable
not taken into

• Statistic
cal—that is, n
rienced report
and jump to t
called their st
tween two lar

25030012485

Source: https://www.industrydocuments.ucsf.edu/docs/anbj0000

average, for every 1,000 trials with no beneficial effects! Zelen in fact has said, "We may now have reached an impasse in cancer chemotherapy in which there are large numbers of false-positive therapies in the clinic,"[3] leading physicians down many false paths.

Amazingly, most false positives probably remain undetected. Scientists do not profit much professionally by reporting negative results. Journal editors are not keen on publishing them. Nor are scientists keen on doing costly and time-consuming studies that merely confirm someone else's work, so "confirmatory studies are rare," Zelen reports.

• Statistical significance alone does not mean there is a cause and effect. *Correlation* or *association* is not causation. Remember the rooster who thought his crowing made the sun rise? Unless an association is so powerful and so constantly repeated that the case is overwhelming, association is only a clue, meaning more study or confirmation is needed.

To statisticians, incidentally, there is this important difference between correlation and association: *Association* means there is at least a possible relation between two variables. A *correlation* is a measure of the association.

• If the number of subjects is too small, an unimpressive *P* value may simply mean that there were too few subjects to detect something that might have shown an effect in more subjects. Highly "significant" *P* values can sometimes adorn negligible differences in large samples.

• An impressive *P* value might also be explained by some other variable or variables—other conditions or associations—not taken into account.

• Statistical significance does not mean biological, clinical—that is, medical—or practical significance, though inexperienced reporters sometimes see or hear the word "significant" and jump to that conclusion, even reporting that the scientists called their study "significant." Example: A tiny difference between two large groups in mean hemoglobin concentration, or

2503012486

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

red blood count (say, 0.1 g/100 mL, or a tenth of a gram per 100 milliliters), may be statistically significant yet medically meaningless.[4]

• Eager scientists can consciously or unconsciously manipulate the *P* value by failing to adjust for other factors, by choosing to compare different end points in a study (say, condition on leaving the hospital rather than length of survival), or by choosing the way the *P* value is calculated or reported.

There are several mathematical paths to a *P* value, such as the chi-square (χ^2), *t*, *F*, *z*, and paired *t* tests. All may be legitimate. But be warned. Dr. David Salsburg of Pfizer, Inc., has written in the *American Statistician* of the unscrupulous practitioner who "engages in a ritual known as 'hunting for *P* values'" and finds ways to modify the original data to "produce a rich collection of small *P* values" even if those that result from simply comparing two treatments "never reach the magical .05."[5]

"If you look hard enough through your data," contributes an investigator at a major medical center, "if you do enough subset analyses, if you go through 20 subsets, you can find one"—say, "the effect of chemotherapy on premenopausal women with two to five lymph nodes"—"with a *P* value less than .05. And people do this."

"Statistical tests provide a basis for probability statements," writes Dr. John Bailar, "only when the hypothesis is fully developed before the data are examined. . . . If even the briefest glance at a study's results moves the investigator to consider a hypothesis not formulated before the study was started, that glance destroys the probability value of the evidence at hand." (At the same time, Bailar adds, "review of data for unexpected clues . . . can be an immensely fruitful source of ideas" for new hypotheses "that can be tested in the correct way." And occasionally "findings may be so striking that independent confirmation . . . is superfluous.")[6]

A rather sophisticated—and possibly touchy—line of questioning that some reporters might want to try if they're skeptical: *How did you arrive at your* P *value? Did you use the test planned in*

advance in your pro:
report the best-soun.
And you m.

The laws of
even impossible-
We've all tal
and bumped int
don't know, but
work, the chanc
1,024. Yet I wo
year period. W.
statisticians call
few people with
cover, there will
birth defects tha
in a great while
In a large
unusual. They
and ofte
duce uni
evidence. "The
large number
occurred. They
ity are wrong."
"We [repc
dence," Philip
and we are rig
mind our reac
from a few in
member. The
A statistic
people or, a st
whom such ar
The chance of
oping leukemi

2503012487

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

advance in your protocol or study design, or did you apply several tests, then report the best-sounding one?

And you may think of other questions.

The laws of probability also teach us to *expect* some unusual, even impossible-sounding events.

We've all taken a trip to New York or London or someplace and bumped into someone from home. The chance of that? I don't know, but if you and I tossed for a drink every day after work, the chance that I would ever win 10 times in a row is 1 in 1,024. Yet I would probably do so sometime in a four- or five-year period. What I like to call the Law of Unusual Events—statisticians call it the Law of Small Probabilities—tells us that a few people with apparently fatal illnesses will inexplicably recover, there will be some amazing clusters of cases of cancer or birth defects that will have no common cause, and I may once in a great while bump into a friend far from home.

In a large enough population such coincidences are not unusual. They are the rule. They produce striking anecdotes and often striking news stories. In the medical world they produce unreliable, though often cited, testimonial or anecdotal evidence. "The world is large," Vogt notes, "and one can find a large number of people to whom the most bizarre events have occurred. They all have personal explanations. The vast majority are wrong."[7]

"We [reporters] are overly susceptible to anecdotal evidence," Philip Meyer writes. "Anecdotes make good reading, and we are right to use them. . . . But we often forget to remind our readers—and ourselves—of the folly of generalizing from a few interesting cases. . . . The statistic is hard to remember. The success stories are not."[8]

A statistic to ask about is the *denominator*—the number of people or, a statistician would say, the *population* or domain—in whom such an event might happen. Zelen cites this example: The chance of any youngster between ages five and nine developing leukemia is 3 in 100,000 per year. In a school with 100

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

2503012488

children of this age group, we would expect only 3 cases in 100 years. But in this nation with thousands of schools, we would occasionally—such is chance—find schools with 3 or more cases in a single year. "Then one is faced with the problem of interpretation," Zelen says. "Is this one of those rare events that is surely going to be observed? Or is it due to some causal factor?"

A reporter in this instance might ask a statistician at the National Cancer Institute or a medical center, What is the chance of such an event in such a population? How many similar unusual events are probably never reported?

"Power" and Numbers

This gets us to another statistical concept: *power.* Statistically, "power" means the probability of finding something if it's there. Example: Given that there is a true effect, say a difference between two medical treatments or an increase in cancer caused by a toxin in a group of workers, how likely are we to find it?

Sample size confers power. Statisticians say, "Funny things can happen in small samples without meaning very much" . . . "There is no probability until the sample size is there" . . . "Large numbers confer power" . . . "Large numbers at least make us sit up and take notice."*

All this concern about sample size can also be expressed as the *law of large numbers,* which says that as the number of cases increases, the probable truth of a conclusion or forecast increases. The *validity* (truth or accuracy) and *reliability* (reproducibility) of the statistics begin to converge on the truth.

We already learned this when we talked about probability.

*There is another unrelated use of the word "power." Scientists commonly speak of increasing or "raising" some quantity *by a power of* 2 or 3 or 100 or whatever. "Power" here means the product you get when you multiply a number by itself one or more times. Thus, in $2 \times 2 = 4$, 4 is the second power of 2, or to put it another way, there are two 2's in your equation. This is commonly written 2^2 and known as 2 to the second power or just 2 to the second. In $2 \times 2 \times 2 = 8$, 2 has been raised to the third power. When you think about 2^{100}, you see the need for the shorthand.

2503012489

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

But by thinking of power as statisticians do—as a function of both sample size and the accuracy of measurement, since that too affects the probability of finding something—we can see that if the number of treated patients is small in a medical study, a shift from success to failure in only a few patients could dramatically decrease the success rate.

If six patients have been treated with a 50 percent success rate, the shift to the failure column of just one would cut the success rate to 33 percent. And the total number is so small in any case that the result has little reliability. The result might be valid or accurate, but it would not be generalizable—it would not have reliability until confirmed by careful studies in larger samples. The larger the sample, and assuming there have been no fatal biases or other flaws, the more confidence a statistician would have in the result.

One canny science reporter, Lewis Cope, says,

> I have my own "rule of two." If someone makes some numerical claim, I look at the numbers, then see how much I might change the finding by adding or subtracting two from any of the figures. For example, someone says there are five cases of cancer in a community. Would it seem meaningful if there were three?
>
> Or if there were eight cases this year but four the year before—a 100 percent increase—I ask myself, "If I add two cases to last year's total and subtract two from this year's, is there a chance things haven't changed, except by chance?" This approach will never supplant refined analysis. But by playing around with the numbers this way—I sometimes try three instead of two—a reporter can often spot a potential problem or error.

A statistician says, "This can help with small numbers but not large ones." Mosteller contributes "a little trick I use a lot on counts of any size." He explains, "Let's say some political unit has 10,000 crimes or deaths or accidents this year. Has something new happened? The minimum standard deviation [see

2503012490

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

page 33] for a number like that is 100—that is, the square root of the original number. That means the number may vary by a minimum of 200 every year without even considering growth, the business cycle, or any other effect. This will supplement your reporter's approach."

Looking for error in reported results, statisticians try to spot both false positives and false negatives. The *false positive* (or *Type I* or *alpha error* in statistical language you may see) is to find a result or effect where there is none. The *false negative* (or *Type II* or *beta error*) is to miss an effect where there is one. The latter is particularly common when there are small numbers. "There are some very well conducted studies with small numbers, even five patients, in which the results are so clear-cut that you don't have to worry about power," says Dr. Relman. "You still have to worry about applicability to a larger population, but you don't have to doubt that there was an effect. When results are negative, however, you have to ask, How large would the effect have to be to be discovered?"

Many scientific and medical studies are underpowered—that is, they include too few cases. "Whenever you see a negative result," another scientist says, "you should ask, What is the power? What was the chance of finding the result if there was one?" One study found that an astonishing 70 percent of 71 well-regarded clinical trials that reported no effect had too few patients to show a 25 percent difference in outcome. Half of the trials could not have detected a 50 percent difference.[9]

A statistician scanned an article on colon cancer in a leading journal. "If you read the article carefully," he said, "you will see that if one treatment was better than the other—if it would increase median survival by 50 percent, from five to seven and a half years, say—they had only a 60 percent chance of finding it out. That's little better than tossing a coin!"

The weak power of that study would be expressed numerically as .6, or 60 percent. Scan an article's fine print or footnotes, and you will sometimes find such a *power statement*. Most

25030124 91

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

authors still don't report one, but the practice is growing, especially when results are negative.

How large is a large enough sample? One statistician calculated that a trial has to have 50 patients before there is even a 30 percent chance of finding a 50 percent difference in results.

Sometimes large populations indeed are needed.[10] If some kind of cancer usually strikes 3 people per 2,000, and you suspect that the rate is quadrupled in people exposed to substance X, you would have to study 4,000 people for the observed excess rate to have a 95 percent chance of reaching statistical significance. The likelihood that a 30-to-39-year-old woman will suffer a myocardial infarction, or heart attack, while taking an oral contraceptive is about 1 in 18,000 per year. To be 95 percent sure of observing at least one such event in a one-year trial, you would have to observe nearly 54,000 women.[11]

Even the lack of an effect—statistically sometimes called a zero numerator—can be a trap. Say, someone reports, "We have treated 14 leukemic boys for five years with no resulting testicular dysfunction"—that is, zero abnormalities in 14. The question remains, how many cases would they have had to treat to have any real chance of seeing an effect? The probability of an effect may be small yet highly important to know about.

All this means you must often ask, *What's your denominator? What's the size of your population?** A disease rate of 10 percent in 20 individuals may not mean much. A 10 percent rate in 200 persons would be more impressive. A rate is only a figure. Always try to get both the numerator and the denominator.

The most important rule of all about any numbers: Ask for them. When anyone makes an assertion that should include numbers and fails to give them, when anyone says that most people, or even X percent, do such and such, you should ask,

*And know that to a statistician a population does not necessarily mean a group of people. Statistically, a *population* is any group or collection of pertinent units—units with one or more pertinent characteristics in common—people, events, objects, records, test scores, or physiological values (like blood pressure readings). Statisticians also use the term *universe* for a whole group of people or units under study.

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

2503012492

What are your numbers? After all, some researchers reportedly announced a new treatment for a disease of chickens by saying, "33.3 percent were cured, 33.3 percent died, and the other one got away."

Bias and Confounders

One scientist once said that lefties are overrepresented among baseball's heavy hitters. He saw this as "a possible result of their hemispheric lateralization, the relative roles of the two sides of the brain." A critic who had seen more ball games said some simpler covariables could explain the difference. When they swing, left-handed hitters are already on the move toward first base. And most pitchers are right-handers who throw most often to right-handed hitters.[12]

Scientist A was apparently guilty of *bias,* meaning the introduction of spurious associations and error by failing to consider other influential factors. The other factors may be called *covariables, covariates, intervening* or *contributing variables, confounding variables,* or *confounders.* A simpler term may be "other explanations."

Statisticians call bias "the most serious and pervasive problem in the interpretation of data from clinical trials" . . . "the central issue of epidemiological research" . . . "the most common cause of unreliable data." Able and conscientious scientists try to eliminate biases or account for them in some way. But not everybody who makes a scientific, medical, or environmental claim is that skilled. Or that honest. Or that all-powerful. Some biases are unavoidable by the very difficulty of much research, and the most insidious biases of all, says one statistician, are "those we don't know exist."

Some biases may be uncovered by assiduous investigation. A father noticed that every time one of his 11 kids dropped a piece of bread on the floor, it landed with the buttered side up. "This utterly defies the laws of chance," he exclaimed. Close examination disclosed the cause: The kids were buttering their bread on both sides.

25030124 93

Source: https://www.industrydocuments.ucsf.edu/docs/gnbj0000

I told this story to one statistician, who said, "I was once called about a person who had won first, second, and third prizes in a church lottery. I was asked to assess the probability that this could have happened. I found out that the winner had bought nearly all the tickets."

He had of course asked the obvious question for both scientist and reporters: *Could the relationship described be explained by other factors?*

Not everyone will tell you, of course, for bias is a pervasive human failing. As one candid scientist is said to have admitted, "I wouldn't have seen it if I hadn't believed it." Enthusiastic investigators often tell us their findings are exciting. But they may be so exciting that the investigators paint the results in over-rosy hues.

Other powerful human drives—the race for academic promotion and prestige, financial connections—can also create conscious or unconscious conflicts of interest or attitudes that feed bias. Dr. Thomas Chalmers of Mount Sinai Medical Center in New York tells of a drug trial, financed by a pharmaceutical firm, in which both the head of the study committee and the main statisticians and analysts were the firm's employees, though not so identified in any credits. He tells of a study of oral drugs for diabetes in which the fact that the first author had previously published 14 articles on the subject, and in 7 had acknowledged support by the drug manufacturers, was "not known to the reader."

In contrast, Chalmers describes a study also financed by a drug firm but with a contract specifying a study protocol designed by independent investigators and monitored by an outside board less likely to be influenced by a desire for a favorable outcome. "It is never possible to eliminate" potential conflicts of interest in biomedical research, he concludes, but they should be disclosed so others can evaluate them.[13]

Even a genius may be biased. Horace Freeland Judson of Johns Hopkins University tells how Isaac Newton experimented with prisms and lenses and developed a theory of color, light,

2503012494

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

and the solar spectrum. He did not report seeing some dark lines—absorption lines, which mark varying wavelengths—that his instruments must have shown. A modern scientist argues that Newton's theory, not his instruments, had no place for that evidence: "To the observing scientist, hypothesis is both friend and enemy."[14]

For years technicians making blood counts were guided by textbooks that told them two or more "properly" studied samples from the same blood should not vary beyond narrow "allowable" limits. Reported counts always stayed inside those limits. A Mayo Clinic statistician rechecked and found that at least two thirds of the time the discrepancies exceeded the supposed limits. The technicians had been seeing what they had been told to expect and discounting any differences as mistakes. This also saved them from the additional labor of doing still more counting.

Both the *biased observer* and the *biased subject* are common in medicine. A researcher who wants to see a treatment result may see one. A patient may report one out of eagerness to please the researcher. There is also the powerful *placebo effect*. Summarizing many studies, one scientist found that half the patients with headaches or seasickness—and a third of those suffering from coughs, mood changes, anxiety, the common cold, and even the disabling chest pains of angina pectoris—reported relief from a "nothing pill."[15] A placebo is not truly a nothing pill; the mere expectation of relief seems to trigger important effects within the body. But in a careful study the placebo should not do as well as a test medication; otherwise the test medication is no better than a placebo.

Sampling bias is the bugaboo of both political polls and medical studies. Say you want to know what proportion of the populace has heart disease, so you stand on a corner and ask people as they pass. Your sample is biased, if only because it leaves out those too disabled to get around. Your problem, a statistician would say, is *selection*. A political pollster who fails to build a valid probability sample, easy when questioning only a thousand or

2503012495

Source: https://www.industrydocuments.ucsf.edu/docs/gnbj0000

so people from coast to coast, has equally poor selection.[16]

A doctor in a clinic or hospital with an unrepresentative patient population—healthier or sicker or richer or poorer than average—may report results that do not represent the population as a whole. Veterans Administration hospitals, for example, treat relatively few women; their conclusions may apply only to the disproportionate number of lower-income men who typically seek out the VA hospitals' free care. A celebrated Mayo or Cleveland or Ochsner clinic sees both a disproportionate number of difficult cases and a disproportionate number of patients affluent and well enough to travel. The famed Kinsey reports were valuable revelations of sexual behavior but flawed because the samples consisted disproportionately of upper middle-class men and women and of those willing to talk.

An investigator may also introduce bias by *constraining*, or distorting, a sample—by failing to reveal *nonresponse* or by otherwise "throwing away data." A surgeon cites his success rate in those discharged from the hospital after an operation but omits those who died during or just after the procedure. Many people drop out of studies—sometimes they just quit—or they are dropped for various reasons: They could not be evaluated, they came down with some "irrelevant" disorders, they moved away, they died. In fact, many of those not counted may have had unfavorable outcomes had they stayed in the study.

Mosteller tells of a nationwide study of a possibly dangerous anesthetic. The investigators relied on autopsy results at 38 hospitals. Unfortunately, only about 60 percent of the relevant dead had been autopsied, and "anything could have been explained by the missing 40 percent, so that part of the study wound up with a handful of nothing."

The presence of significant nonresponse can often be detected, when reading medical papers, by counting the number of patients treated versus the number of untreated or differently treated *controls*—patients with whom the treated patients are compared. If the number of controls is strikingly greater in a randomized clinical trial (though not necessarily in an epidemio-

2503012496

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

logical or environmental study), there were probably many dropouts. A well-conducted study should describe and account for them. A study that does not may report a favorable treatment result by ignoring the fate of the dropouts—a confounding variable.

Age, gender, occupation, nationality, race, income, socioeconomic status, health status, and powerful behaviors like smoking are all possible confounding—and frequently ignored—variables. In the 1970s, foes of adding fluoride to city water pointed to crude cancer mortality rates in two groups of 10 U.S. cities. One group had added fluoride to water, the other had not, and from 1950 to 1970 the cancer mortality rate rose faster in the fluoridated cities. The National Cancer Institute pointed out that the two groups were not equal: The difference in cancer deaths was almost entirely explained by differences in age, race, and sex. The age-, race-, and sex-adjusted difference actually showed a small, unexplained lower mortality rate in the fluoridated cities.[17]

If you look carefully at the fate of women taking birth control pills, you find that advancing age and smoking are the two great confounders. You must take both into account to find the greatest clusters of ill effects. Smoking has been an important confounder in studies of industrial contaminants like asbestos, in which, again, the smokers suffer a disproportionate number of ill effects.[18]

A 1947 survey of Chicago lawyers showed that those who had mere high school diplomas before entering legal training earned 6.3 percent more, on the average, than college graduates. The confounder here—the real explanation—was age. In 1947 there were still many older lawyers without college degrees, and they were simply older, on the average, and hence more established.[19]

Occupational studies often confront another seeming paradox: The workers exposed to some possible adverse effect turn out to be healthier than a control group of persons without such exposure. The confounder: the well-known *healthy-worker effect.*

2503012497

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

Workers tend to be healthier and live longer than the population in general.

Some studies of workers in steel mills showed no overall increase in cancer, despite possible exposures to various carcinogens. It took a look at black workers alone to find excess cancer. They commonly worked at the coke ovens, where carcinogens were emitted. This was a case where the population had to be *stratified,* or broken up in some meaningful way, to find the facts. Such findings in blacks often may be falsely ascribed to race or genetics, when the real or at least the most important contributing or ruling variables—to a statistician, the *independent variables*—are occupation and the social and economic plights that put blacks in vulnerable settings. The excess cancer is the *dependent variable,* the result.

"In a two-variable relationship," Dr. Gary Friedman explains, "one is usually considered the independent variable, which affects the other or dependent variable."[20] Take the fact that more people get colds in winter. Here weather is commonly seen as the underlying or independent variable, which affects incidence of the common cold, the dependent variable. Actually, of course, some people, like children in school who are constantly exposed to new viruses, are more vulnerable to colds than others. In the case of these children, then, as in the case of the black workers at the coke ovens, there is often more than one independent variable. Also, some people think that an important underlying reason for the prevalence of colds in winter may be that children are congregated in school, giving colds to each other, thence to their families, thence to their families' coworkers, thence to the coworkers' families, and so on. But cold weather—and home heating?—may still figure, perhaps by drying nasal passages and making them more vulnerable to viruses.

The search for *true variables* is obviously one of the main pursuits of the epidemiologist, or disease detective—or of any physician who wants to know what has affected a patient, or of any student of society who seeks true causes. Like colds, many

2503012498

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

medical conditions, such as heart disease, cancer, and probably mental illness, have multiple contributing factors. Where many known, measurable factors are involved, statisticians can use mathematical techniques—the terms you will see include *multiple regression, multivariate analysis,* and *discriminant analysis* and *factor, cluster, path,* and *two-stage least-squares analyses*—to relate all the variables and try to find which are the truly important predictors. Yet some situations, like the striking decline in U.S. heart disease mortality in recent years, defy such analyses. These years have seen several major changes in American life that may play a role: less smoking among men, consumption of a leaner diet, more recreational exercise (though more sedentary work). Medical care is far better, including the treatment of hypertension, which disposes people to heart disease. Many of these variables cannot be well measured, and the effect of some is debatable, so—a common situation in science—the truth remains uncertain.

Variability

Doctors always say, "Most things are better in the morning," and they're mostly right. Most chronic or recurring conditions wax and wane. We tend to wake up at night when the condition is at its worst. Then, no matter what is done by way of treatment the next day, the odds are that we'll feel better.

This is *regression toward the mean:* the tendency of all values in every field of science—physical, biological, social, and economic—to move toward the average. Tall fathers tend to have shorter sons, and short fathers, taller sons. The students who get the highest grades on an exam tend to get somewhat lower ones the next time. The regression effect is common to all repeated measurements.

Regression is part of an even more basic phenomenon: *variation,* or *variability.* Virtually everything that is measured varies from measurement to measurement. When repeated, every experiment has at least slightly different results. Take a patient's

blood pressur
row, and the
different times
vary greatly.

The imp
also measure
and observer
doctors will r
be grossly dif
heart murme
hearing to de
one time to t
cancer resear
usual regular
too well and
enough varia

Biologica
physiology a
tients, react
differ
lations,
within the s

Every p
each with m
such as heig
and—if we
tion—we m
We can't get
need single

Enter
median, and
some idea
properties,

When
mean or *arith*
number of

25030124 99

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

blood pressure, pulse rate, or blood count several times in a row, and the readings will be somewhat different. Take them at different times of day or on different days, and the readings may vary greatly.

The important reasons? In part, fluctuating physiology, but also measurement errors, the limits of measurement accuracy, and observer variation. Examining the same patient, no two doctors will report exactly the same results, and the results may be grossly different. If six doctors examine a patient with a faint heart murmer, only one or two may have the skill or keen hearing to detect it. Experimental results so typically differ from one time to the next that scientific and medical fakers—a Boston cancer researcher, for example—have been detected by the unusual regularity of their reported results, with numbers agreeing too well and the same results appearing time after time, with not enough variation from patient to patient.

Biological variation is the most important cause of variation in physiology and medicine. Different patients, and the same patients, react differently to the same treatment. Disease rates differ in different parts of the country and among different populations, and—alas, nothing is simple—there is natural variation within the same population.

Every population, after all, is a collection of individuals, each with many characteristics. Each characteristic, or *variable,* such as height, has a *distribution* of values from person to person, and—if we would know something about the whole population—we must have some handy summaries of the distribution. We can't get much out of a list of 10,000 measurements, so we need single values that summarize many measurements.

Enter here the familiar *average* or, more exactly, the *mean, median,* and *mode.* These and a few other measures can give us some idea of the look of the whole and its many measurable properties, or *parameters.*

When most of us speak of an average, we mean simply the *mean* or *arithmetic average,* the sum of all the values divided by the number of values. The mean is no mean tool; it is a good way

25030125OO

Source: https://www.industrydocuments.ucsf.edu/docs/[illegible]

to get a typical number, but it has limitations, especially when there are some extreme values. There is said to be a memorial in a Siberian town to a fictitious Count Smerdlovski, the world's champion at Russian roulette. On the average he won, but his actual record was 73 and 1.[21]

If you look at the average salary in a hospital, you will not know that half the personnel may be working for the minimum wage, while a few hundred persons make $100,000 or more a year. You may learn more here from the *median,* the figure that divides a population into two equal halves. The median can be of value when a group has a few members with extreme values, like the 400-pounder at an obesity clinic whose other patients weigh from 180 to 200 pounds. If he leaves, the patients' mean weight might drop by 10 pounds, but the median might drop just 1 pound.[22]

The most frequently occurring number or value in a distribution is called the *mode.* When the median and the mode are about the same, or even more when mean, median, and mode are roughly equal, you can feel comfortable about knowing the typical value.

You still need to know something about the exceptions, in short, the *dispersion* (or spread or scatter) of the entire distribution. One measure of spread is the *range.* It tells you the lowest and highest values. It might inform you, for example, that the salaries in that hospital range from $10,000 to $250,000.

You can also divide your values into 100 *percentiles,* so you can say someone or something falls into the 10th or 71st percentile, or into *quartiles* (fourths) or *quintiles* (fifths). One useful measure is the *interquartile range,* the interval between the 75th and 25th percentiles—this is the distribution in the middle, which avoids the extreme values at each end. Or you can divide a distribution into *subgroups*—those with incomes from $10,000 to $20,000, for example, or ages 20 to 29, 30 to 39, and so on.

All these values can easily be plotted. With many of the things that scientists, economists, or others measure—IQs, for example, and other test scores—we typically tend to see a famil-

iar, bell-shaped
end, or *tail.* T
19th-century
But you may
clusters, a *bim*

A widely
great deal. Nc
tance from th
range, this ha
how spread o
In what one s
in most sets
being measur
average by n
more than 2
than 2.57 sta

"Once yc
shaped distril
the whole pic
curve
variation of t
the more spr

"There is n
depending on th
differences betwe
number of square
of a population ra
result. As in

Sometimes
mean, being an
standard error or a

All the abo

25030125O1

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

iar, bell-shaped *normal distribution,* high in the middle, low at each end, or *tail.* This is the classic *Gaussian curve,* named after the 19th-century German mathematician Karl Friedrich Gauss. But you may also find that the plot has two or more peaks or clusters, a *bimodal* or *multimodal distribution.*

A widely used number, the *standard deviation,* can reveal a great deal. No matter how it sounds, it is not the average distance from the mean but a more complex figure.* Unlike the range, this handy figure takes full account of every value to tell how spread out things are—how dispersed the measurements. In what one statistician calls a truly remarkable generalization, in most sets of measurement "and without regard to what is being measured" only 1 measurement in 3 will deviate from the average by more than 1 standard deviation, only 1 in 20 by more than 2 standard deviations, and only 1 in 100 by more than 2.57 standard deviations.

"Once you know the standard deviation in a normal, bell-shaped distribution," according to Thomas Louis, "you can draw the whole picture of the data. You can visualize the shape of the curve without even drawing the picture, since the larger the variation of the numbers, the larger the standard deviation and the more spread out the curve—and vice versa."

*There is more than one way to calculate it, and there are several variations, depending on the statistician's purpose. A common one is to add the squares of the differences between each number and the mean, then divide that number by the total number of squares, often referred to as the *variance* (minus 1 if you're looking at a sample of a population rather than the whole population). Then calculate the square root of the result. As in

$$s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n - 1}}$$

Sometimes statisticians calculate the *standard deviation of the mean*—this because the mean, being an average, is less variable than single measurements. Some call this the *standard error* or *standard error of the mean.* As in

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

All the above are measures of dispersion.

25030125502

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

Example: If the average score of all students who take the SAT college entrance test is relatively low and the spread—the standard deviation—relatively large, this creates a very long-tailed, low-humped curve of test scores, ranging, say, from around 300 to 1500. But if the average score of a group of brighter students entering an elite college is high, the standard deviation of the scores will be less and the curve will be high-humped and short-tailed, going from maybe 900 to 1500.

"If *I* just told you the means of two such distributions, you might say they were the same," another scientist says. "But if I reported the means and the standard deviations, you'd know they were different, with a lot more variations in one."

From a human standpoint, variation tells us that it takes more than averages to describe individuals. Biologist Stephen Jay Gould learned in 1982 that he had a serious form of cancer. The literature told him the median survival was only eight months after discovery. Three years later he wrote in *Discover,* "All evolutionary biologists know that means and medians are the abstractions," while variation is "the reality," meaning "half the people will live longer" than eight months.

Since he was young, since his disease had been diagnosed early, and since he would receive the best possible treatment, he decided he had a good chance of being at the far end of the curve. He calculated that the curve must be skewed well to the right, as the left half of the distribution had to be "scrunched up between zero and eight months, but the upper right half [could] extend out for years." He concluded, "I saw no reason why I shouldn't be in that small tail. . . . I would have time to think, to plan and to fight." Also, since he was being placed on an experimental new treatment, he might if fortune smiled "be in the first cohort of a new distribution with . . . a right tail extending to death by natural causes at advanced old age."[23]

Statistics cannot tell us whether fortune will smile, only that such reasoning is sound.

Studie
Good

Why think? Why

Sit down before fa
notion, follow hur
shall learn nothin

This is the part I

THERE is
cally than Alz
the brain. At
team imagina
tients aged 64
theoretically p
solution for c

After 18
that accordin
patients show
held his own.
the work and
do a story ar

25030125O3

Source: https://www.industrydocuments.ucsf.edu/docs/

Studies, Good and Bad 4

Why think? Why not try an experiment?

—John Hunter
18th-century British anatomist

Sit down before fact as a little child, be prepared to give up every preconceived notion, follow humbly wherever and to whatever abysses nature leads, or you shall learn nothing.

—Thomas Henry Huxley

This is the part I always hate.

—A mathematician as he approaches the equal sign (in a Sidney Harris cartoon in *American Scientist*)

THERE is no disease that strikes older people more tragically than Alzheimer's disease, which makes a useless tangle of the brain. At a prestigious New England university a research team imaginatively inserted catheters into the skulls of four patients aged 64 to 73 to deliver a continuous infusion of either a theoretically promising drug or, alternately, an ineffectual saline solution for comparison.

After 18 months the investigators published a paper saying that according to observations by the patients' families, three patients showed marked improvement and the fourth at least held his own. Fascinating, of course. Some reporters learned of the work and began inquiring. The investigators let a TV crew do a story and also held a news conference, with one patient

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

2503012504

brought forth for on-camera testimonials. Except for some newspapers that decided to print nothing, the story flew far and wide.

The head investigator, a chief resident in neurosurgery, cautioned that the results, though encouraging, were "very early" and "certainly do not prove this is an effective treatment." He advised healthy skepticism. But headlines unequivocally read: "Alzheimer's Test Found Successful," "Alzheimer's: A New Promise," "First Breakthrough Against Alzheimer's," "Pump Offers Hope," "Possible Alzheimer's Cure."

Within two months the medical center logged 2,600 phone calls, mainly from desperate families, and critics began asking why a press conference had been held, since a study of only four patients—with unblinded investigators getting their assessments from hopeful families—meant little.

Harvard's Dr. Jay Winsten concluded that "the decision to hold a press conference . . . far outweighed in impact the modulating effect of the investigators' qualifying language. The visual impact of [one] patient's on-camera testimonials all but guaranteed that TV coverage would oversell the research, despite any qualifying language."[1]

When dubious claims are made—about Alzheimer's, a new cancer drug, a possible AIDS cure—and the claims get widely reported, there is commonly a lot of postmortem clucking and soul-searching among reporters and editors. Then someone else makes some sensational claim, and the same thing may happen all over again.

The biggest error in medical science, according to Dr. Thomas Chalmers, is "the uncontrolled pilot study in which the investigators try a treatment on 10 patients, and if it seems to work . . . are tempted to report it" to fellow scientists, let alone the media.[2]

All science is only a stab at the truth. Even with the best of statistics, "We scientists don't know how to tell the whole truth," Mosteller reminds us.[3] Outside this honest limitation lie vast realms of inadequate science with plausible-sounding yet shaky

25030125O5

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

statistics. A French physician, Pierre Charles Alexandre Louis, said 150 years ago, "The only reproach which can be made to the numerical method" is that it "requires much more labor and time than the most distinguished members of our profession" often give it. "Some days," says one modern statistician, "I think every idiot in the country who can put his hands on a computer program thinks he's a statistician."

The big problems of statistics, say its best practitioners, have little to do with computations and formulas. They have to do with judgment, we're told, with how to design a study, how to conduct it, then analyze and interpret the results. In a day of frenzied media competition for the public's eye and ear—and many chances to do harm by shaky reporting—journalism too calls for sophisticated judgment. How, then, can we have some hope of telling which studies seem credible, which we should report?

A fundamental principle is that every conscientiously conducted study has a careful *design:* a method or plan of attack to include the right kind and number of patients or petri dishes and to try to eliminate bias. Different problems require different methods, and one of the most basic questions in science is, *Can this kind of experiment, this design, yield the answer?*

This is not a simple question for a reporter to answer, but there is much we can know. What kinds of studies, what kinds of numbers and controls and methods, should we look for?

Experiments versus Seductive Anecdotes

Students and eggs can be graded, citizens and cities can be credit-rated, and scientific evidence can be weighed according to what has been called a hierarchy of evidence. Some kinds of studies carry little weight, some more, some a great deal.

Science and medicine started with *anecdotes,* unreliable as far as generalization is concerned, yet provocative. Anecdotes matured into systematic *observation,* the most ancient form of science. Observation told the ancients much about the stars, it

2503012506

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

told the pharaohs' physicians much about the sick, and it is still important, for simple "eyeballing" has developed into *data collection* and the recording of *case histories*. These are respectable, yea, indispensable methods yet still only one part of science. Case histories may not be typical, or they may reflect the beholder. Medicine continues to be plagued by Big Authorities who insist, "I know what I see."

There can be useful, even inspired, observation and analysis of *natural experiments*. Excess fluoride in some waters hardened teeth, and this observation led to fluoridation of drinking water to prevent tooth decay. There are also man's inadvertent experiments, disastrous and benign, to be studied. Hiroshima triggered wide analysis of the effects of nuclear radiation, invaluable yet frustrating because there were no good measures of exposure levels, a gap that has caused confusion and controversy ever since.

In 1585 or so, Galileo dropped those weights from a tower and helped invent the *scientific experiment:* a study in which the experimenter *controls* the conditions—controlled conditions are the heart of the experimental method—and records the effect. Experiments on objects, animals, germs, and people matured into the modern *experimental study,* in which the experimenter typically changes only one or some other planned number of variables to see the outcome.

Clinical Trials

The experimental method is the essence of experimental medicine's current "gold standard":[4] the *controlled, randomized clinical trial.* At its best, the investigator tests a treatment or drug or some other intervention by randomly selecting at least two comparable groups, the *experimental group* that is tested or treated and a *control group* that is observed for comparison.

True clinical trials are expensive and difficult. It has been estimated that of 100 scheduled trials, 60 are abandoned, not

25030125O7

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

implemented, or not completed, whether for lack of funds, difficulty in recruiting or keeping patients, toxicity or other problems, or, sometimes, rapid evidence of a difference in effect (making continued denial of effective treatment to a control group unethical). Another 20 trials produce no noteworthy results, and just 20, results worth publishing. Clinical trials nonetheless are called the strongest, most precise, most decisive way to evaluate medical interventions and learn true causation. Randomized clinical trials proved that new drugs could cut the heart attack death rate, that treating hypertension could prevent strokes, and that polio, measles, and hepatitis vaccines worked. No doctor, observing a limited number of patients, could have shown these things.

Types of clinical studies include the following:

• Among the most reliable are *parallel studies* comparing similar groups given different treatments, or a treatment versus no treatment. But such studies are not always possible.

• In *crossover studies* the same patients get two or more treatments in succession and act as their own controls. Similarly, *self-controlled studies* evaluate an experimental treatment by control observations during periods of no treatment or of some standard treatment. There are pitfalls here. Treatment A might affect the outcome of treatment B, despite the usual use of a *washout period* between study periods. Patients become acclimated: They may become more tolerant of pain or side effects or, now more health-conscious, may change their ways. The controls—the patients in a control group—don't always behave in parallel studies either: In one large-scale trial of methods to lower blood cholesterol and risk of heart disease, many controls adopted some of the same methods—quitting cigarette smoking, eating fewer fats—and reduced their risk too.

• Investigators often use *historical controls* (meaning comparison with old records: historically the cure rate has been 30 percent, say, and the new therapy cures 60 percent) or other *external controls* (such as comparison with other studies). These

2503012508

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

controls are often misleading—the groups compared are frequently not comparable, the treatments may have been given by different methods—but they are still at times useful.

What Makes a Study Honest?

Obviously, all studies, including the best, have potential pitfalls:

• *Lack of adequate controls* is fatal if you really want to put the results in the bank.

• The *group or sample studied,* 10 people or 10,000, must be *large* enough to get a valid result and *representative* enough to apply to a larger population. Because people vary so widely in their reactions, and a few patients can fool you, fair-sized groups of patients are usually needed. And enough of the right kind of subjects are needed for a suitable sample. Picking patients for a medical study is no different from picking citizens to be questioned in a political poll. In both, a sample is studied, and *inferences*—the outcome of an election, the results in patients in general—are made for a larger population.

To get a large enough sample, medical researchers more and more try to conduct *multicenter trials,* which are appealing because they can include hundreds of patients, but expensive and tricky because one must try to maintain similar patient selection and quality control at 10 or 100 institutions. Successful multicenter trials established the value of controlling hypertension to prevent strokes. They demonstrated the strong probability that less extensive surgery is as effective as more drastic surgery for many breast cancers.

• The *sample should be randomized*—divided by some random method into comparable experimental and control groups. Randomization can easily be violated. A doctor assigning patients to treatment A or B may, seeing a particular type of patient, say or think, "This patient will be better on B."

If treatment B has been established as better than A, there should be no random study in the first place and certainly no

25030125O9

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

study of that doctor's patient. When randomization is violated, "the trial's guarantee of lack of bias goes down the drain," says one critique. As a result, patients who consent to randomization are often assigned to study groups according to a list of computer-generated random numbers.

• *To combat bias*—the influence of confounding variables—and get answers applicable to various populations, the sample or study population must often be *stratified*, or separated into groups by age, sex, socioeconomic status, and so on. Failure to stratify can hide true associations. The role of high-absorbency tampons in toxic shock syndrome was clarified only when the cases were broken down by precise type of tampon used.

The identification of important subcategories of patients can be tricky indeed. A study of open-heart surgery patients may fail to separate out those who had to wait for their surgery. But some patients die waiting, and those left are relatively stronger patients who do better, on the average, than those treated immediately after diagnosis.

We reporters may also fail to pay attention to stratification, or distribution. In early 1985 the President's Council of Economic Advisers reported that—to quote the page-one lead in a major newspaper—"elderly Americans have achieved economic parity with the rest of the population and no longer are a disadvantaged group." Not for several paragraphs, now on an inside page, did the story note that "there's a lot of variability," and older people are also "more likely . . . to have members with incomes below the average of their age group."[5] In short, there are still many elderly trapped in poverty.

• *To combat bias in investigators or patients,* studies should be *blinded*—to the extent feasible, *single-*, *double-*, or, best of all, *triple-blinded,* so that neither the doctors nor the nurses administering a treatment nor the patients nor those who assess the results know whether today's pill is treatment A, treatment B, or an ineffective *placebo.* Otherwise, a doctor or patient who yearns for a good result may see or feel one when the "right" drug is given. There is a tale of an overzealous receptionist who, knowing

25030125 10

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

which patients were getting the real drug and not the placebo, was so encouraging to these patients that they began saying they felt good, willy-nilly.[6]

Barring observant receptionists, the use of a placebo—from the Latin meaning "I shall please"—may help maintain blindness. Placebos actually give some relief in a third of all patients, on the average, in various conditions. The effect is usually temporary, however, and a truly effective drug ought to work substantially better than the placebo.

Blinding is often impossible or unwise. Some treatments don't lend themselves to it, and some drugs quickly reveal themselves by various effects. But an unblinded test is a weaker test.

• Finally, what makes a study honest is honesty. John Bailar warns of deliberate or careless deceptions that seem to be universally accepted today, practices that sometimes have much value but at other times are "inappropriate and improper and, to the extent that they are deceptive, unethical." Among them: the selective reporting of findings, leaving out some that might not fit the conclusion; the reporting of a single study in multiple fragments, when the whole might not sound so good; and the failure to report the low power of some studies, their inability to detect a result even if one existed.[7]

Dr. Charles Moertel of the Mayo Clinic says,

> Probably the majority of cancer patients treated with chemotherapy today are receiving regimens that have not been proved effective by randomized trial. . . . Many articles published in our major journals make claims for fantastic therapeutic accomplishments with no randomized controls. . . . Many, if not most, of the randomized studies . . . are of such poor quality that their results are unbelievable. . . . Precious few have withstood the scrutiny of carefully designed confirmatory scientific study.

He calls a multitude of poor methods statistical legerdemain: "the games we play, trying to squeeze out that little bit of breakthrough." Why the pressure to play them? "Salvation," Dr.

25030125 11

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

David Salsburg answers. "Fruit in this world (increases in salary, prestige, invitations to speak) and beyond this life (continual references in the citation index)."[8]

Epidemiology: Hippocrates to AIDS

Clinical studies deal with patients. Epidemiology deals with populations, which sometimes are large groups of patients. Epidemiology seeks the causes of both health and disease by placing a population under its own kind of microscope, the *epidemiological investigation.*

Epidemiological studies in many ways parallel clinical studies—some studies are both—and are subject to many of the same pitfalls and rules, like avoiding bias and stratifying to get the right answers about the right subgroups. An old saw, in fact, goes, an epidemiologist is a physician broken down by age and sex.

Epidemiology in its early days was concerned wholly with epidemics of typhoid, smallpox, and other infections. But epidemiologists today also ask, "What should we eat and how should we live to stay healthy?" and they study large groups to see how the healthiest and unhealthiest live. Hippocrates has been called the first environmentalist because he observed that it was healthier to live in high places than in low ones. Anticipating today's environmentalists, he blamed bad air and bad water and may have been partly right. But he failed to stratify; otherwise he might have noticed that the people who lived high were also wealthier and better nourished than those who lived low.[9]

In 1740 Percival Pott scored a famous epidemiological success by observing the high rate of scrotum cancer in London's chimney sweeps and correctly blaming it on their exposure to soot—burned organic material, much like a smoked cigarette. A century later, John Snow, plotting London cholera cases on a map and noting a cluster around one source of drinking water, removed the handle from the now famed Broad Street pump and helped end a deadly epidemic. The 19th-

2503012512

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

century French advocate of statistical methods, Pierre Louis, observed hospital patients and helped stop the use of bleeding as a treatment. Ignaz Semmelweis showed that doctors' dirty hands transmitted deadly childbed fever to mothers.

Modern epidemiologists successfully indicted smoking as a cause of lung cancer and heart disease and identified the association of fats and cholesterol with clogging of the arteries. They evaluate vaccines, assess new methods of health care delivery, and track down the causes of new scourges like AIDS, toxic shock syndrome, and Legionnaires' disease, all by several methods. All are valuable. All are full of traps.

• Epidemiology, like all of science, started with *observational* studies, and these remain important. They are weak and uncertain, we have noted, when it comes to determining cause and effect. Yet observation is how we first learned of the unfortunate effects of toxic rain, Agent Orange, cigarette smoking, and many sometimes helpful, sometimes harmful medications—and of certain sexual practices and addicts' use of dirty needles on AIDS.

• Some observational studies are simply *descriptive*—describing the incidence, prevalence, and mortality rates of various diseases, for example. Other, *analytic* studies seek to analyze or explain: the Seven-Country Study, for example, that helped associate high meat and dairy fat and cholesterol consumption with excess risk of coronary heart disease. *Ecological* studies look for links between environmental conditions and illness. Human migrations—like that of the Japanese who come to the United States, eat more fat, and get more disease than they did in Japan—are among valuable *natural experiments*.

• The simplest observational measurement is a *count*. *Sampling* is just a more sophisticated kind of count. You can't count or question everybody, so you seek a sample that represents the whole. Many epidemiological *surveys* rely on samples—among them, government surveys of health and nutritional habits. Samples and surveys often use *questionnaires* to get information.

A sample or survey is never more than a snapshot of the

25030125 13

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

scene at the moment; it can't portray an ever-changing picture unless frequently repeated. Questionnaires may be no better than the quality of the answers, written or verbal. One survey compared patients' reporting of their current chronic illnesses with those their doctors recorded. The patients failed to mention almost half of the conditions the doctors detected over the course of a year. And whether it comes to illness, diets, or drinking, people tend to put themselves in the best possible light. They often say both yes and no to the same question in different form. A survey may stand or fall on the use of sophisticated ways to get accurate information.

• Epidemiologists' studies may also be *prevalence studies, case-control studies,* or *cohort studies.* A *prevalence study,* also called a *current* or *cross-sectional study* is a wide-angle snapshot of a population: a look at the rate of disease X or at toxic agent X and its possible effects by age, sex, or other variables. A political poll is such a study: A cross section of the nation is examined in a period of a few days.

A *case-control study* examines *cases* and *controls* for a close-up of a disease's relationship to other factors in a small, intensively examined group. The nation hears of cases of toxic shock syndrome, mainly in young women. The federal Centers for Disease Control launches a *field investigation* to find a series of patients, or *cases,* confirm the diagnosis, then interview them and their families and other contacts to assemble careful case histories that cover, hopefully, all possible causes or associations. This group is then compared with a randomly selected but matched *compeer group,* or *control group,* of healthy young women of like age and other characteristics.

The results need to be interpreted with great caution, but the case-control study is often a quick, highly useful and relatively easy, low-cost first approach or fishing expedition to assemble clues about causes or even a working hypothesis. Or it may test some hypothesis. A case-control study pinpointed the use of tampons (later found to be certain high-absorbency ones) as the main villain in toxic shock. The relationship of cigarette

2503012514

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

smoking to lung cancer, the association of birth control pills with blood vessel problems, and the transmission patterns of AIDS were identified in case-control studies that pointed to the need for broader investigation.

Cohort or *incidence studies* are motion pictures. They pick a group of people, or *cohort*—a cohort was a unit of a Roman legion—often stratify or divide them into subgroups, then follow them over time, often for years, to see how some disease or diseases develop. These studies are costly and difficult. Subjects drop out or disappear. Large numbers must be studied to see rare events. But cohort studies can be powerful instruments and substitutes for randomized experiments that would be ethically impossible. You can't ethically expose a group to an agent that you suspect would cause a disease. You can watch a group so exposed.

The noted Framingham study of ways of life that might be associated with developing heart disease has followed more than 5,000 residents of that Massachusetts town since 1948. The American Cancer Society's 1952–55 study of 187,783 men aged 50 to 69, with 11,780 of them dying during that period, did much to establish that cigarette smoking was strongly associated with developing lung cancer.[10]

• Many epidemiological, as well as clinical, studies are handicapped because they must be *retrospective*. They look back in time—at medical records, vital statistics, or people's recollections (for example, those collected in interviews in a case-control study). People who have a disease are questioned to try to find common habits or exposures. Women with cervical cancer are interviewed to see how many took possibly guilty hormones and how many did not. People who live around a Love Canal are asked if they have been ill.

Retrospective studies are notoriously unreliable. Memories fail or play tricks. Old records are poor and misleading. Definitions of diseases and methods of diagnosis vary sharply over the years. The patients you find may not be representative. A retrospective study, however intriguing, generally only says that there may be something here that ought to be investigated.

25030125 15

Source: https://www.industrydocuments.ucsf.edu/docs/gnbj0000

(There are exceptions. Dr. Gary Friedman writes, "A retrospective study can be quite reliable if based on data carefully collected in the past. A revealing study of mortality in radiologists was a retrospective cohort study based on good data.")

• A *prospective study,* in contrast—like the Framingham and the American Cancer Society studies—looks forward. It focuses sharply on a selected group who are all followed by the same statistical and medical techniques. Dr. Eugene Robin at Stanford tells how four separate retrospective clinical studies affirmed the accuracy of a test for blood clots in the lungs. When an adequate prospective clinical trial was done, most of the backward looks were proved wrong.[11]

• Epidemiology also includes *experimental studies,* the classical experiments of science on a larger human scale. These are typically *intervention studies.* There is some intervention or manipulation; something is done to some of the subjects.

The massive and hugely successful 1954 field trial of the Salk polio vaccine was a classic intervention trial and a clinical trial too, with 401,974 first- to third-graders assigned at random to either a vaccinated group or a control group injected with a placebo, or dummy shot—and another 947,171 children divided between vaccinated second-graders and unvaccinated first- and third-graders acting as controls. In addition, in all participating states or counties, the investigators studied and counted all cases of polio in a grand total of 1,829,916 children: those who had taken part in the study and those who had not. In the placebo areas, the study was also triple-blinded: neither the vaccinators, the subjects, nor the doctors who examined the subjects later for polio knew which children got which kind of shot.[12]

Another successful intervention study, a *community trial,* established the value of fluoridating water supplies to prevent tooth decay. Some towns had their water fluoridated; some did not. Blinding was impossible, but the striking difference in dental caries that resulted could not have been caused by any placebo effect.

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

2503012516

Questions Reporters Can Ask 5

Just because Dr. Famous or Dr. Bigshot says this is what he found doesn't mean it is necessarily so.

—Dr. Arnold Relman

Ask to see the numbers, not just the pretty colors.

—Dr. Richard Margolin
National Institutes of Health,
describing PET scans to reporters

WHAT questions should we reporters ask—to make our news solid, to report the more valid claims and ignore the weak and phony? When a scientist or physician or anyone else says, "I've discovered that . . . ," what should we ask?

In 1949, a year after Britain's National Health Service—"socialized medicine"—was launched, my editors sent me to Britain to see how it was working. A bit stumped, I asked Dr. Morris Fishbein, the provocative genius who long edited the *Journal of the American Medical Association,* "How can I, a reporter, tell whether a doctor is doing a good job?" He immediately said, "Ask him how often he has a patient take off his shirt."

His lesson was plain: No physical examination is complete unless the patient takes off his or her clothes. Most reporters are not skilled statisticians, but we can ask some similarly revealing questions. Many of these are not even statistical, just simple ones that, like Fishbein's, probe soft spots and often disclose either a conscientious approach or one that can't be trusted.

We can learn here from one method of science. We said

earlier that a pr
seeking truth, of
A is no better tha
sees whether or
much like the la
cutor to prove
guilty. A reporte
should be equal
words or thoug
If an inves
case, you may
since a good sci
for you. The
something.
Here are s
ple and obviou
want to ask th

How do yo
ment? Wh
Answer
"I've seen
block . . ." m
gation, may b
anything like

What kin
design? And a
What wa
case-control, pro
ter for kinds
people just s
conclusion w
medical edito
study? What
answer?

25030125 17

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

earlier that a properly skeptical scientist, starting a study and seeking truth, often begins with a *null hypothesis*—that treatment A is no better than treatment B, that there's nothing there—then sees whether or not the evidence disproves it. This approach is much like the law's presumption of innocence: It is for the prosecutor to prove beyond reasonable doubt that the suspect is guilty. A reporter, without being cynical and believing nothing, should be equally skeptical and greet every claim by saying, in words or thought, "Show me."

If an investigator or claimant is competent and has a good case, you may have to ask none or very few of these questions, since a good scientific presentation should answer most of them for you. The need for a lot of questions could itself tell you something.

Here are some possible questions, then, some of them simple and obvious ones, a few more technical for those who might want to ask them.

How do you know? Have you done a study? Was there an experiment? What is the evidence? Or is the approach just anecdotal? Answers like "In my experience . . . ," "In my hands . . . ," "I've seen 20 cases . . . ," and "There are four cases in our block . . . " may be interesting, may be worth scientific investigation, may be worth a cautious news story, but there is not yet anything like certainty.

What kind of study was it? Was there a systematic research plan or design? And a protocol or set of rules?

What was the study design or method: observational, experimental, case-control, prospective, retrospective, or what? (See the previous chapter for kinds of studies and their uses and limits.) "A lot of people just scrounge around and try to come up with some conclusion without any real plan or design at the start," one medical editor reports. *Was the design drawn before you started your study? What specific questions or hypotheses did you set out to test or answer?*

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

2503012518

Why did you do it that way? Do you think it was the right kind of study to get the answer to this question or problem?

Was it a true human experiment, if possible, with comparable groups picked at random for comparison? If not, why not? And what was the substitute?

If an investigator patiently—you hope—tells you about an acceptable-sounding design, that's worth a brownie point. If the answer is "Huh?" or a nasty one, that may tell you something else.

Are you presenting preliminary data or something fairly conclusive? Are you presenting a conclusion or a hypothesis for further study? "Preliminary" and "interesting" can mean "unproved."

If the result is not reasonably conclusive, should there be further studies and what kind?

How many subjects, patients, cases, or people are you talking about? Are these numbers large enough, statistically rigorous enough, to get the answers you want? Was there an adequate number of patients to show a difference between treatments? Why are you calling a press conference to report on four patients?

Small numbers can sometimes carry weight. And they may sometimes be the only ones possible. "Sometimes small samples are the best we can do," one researcher says. But larger numbers are always more likely to pass statistical muster.

The number studied can also depend on the subject. A thorough physiological study of five cases of some difficult disorder may be important. One new case of smallpox would be a shocker in a world in which smallpox has supposedly been eliminated. In June 1981 the federal Centers for Disease Control reported that five young men, all active homosexuals, had been treated for *Pneumocystis carinii* pneumonia at three Los Angeles hospitals.[1] This alerted the world to what soon became the AIDS epidemic.

Who were your subjects? How were they selected? What were your criteria for admission to the study? Were rigorous laboratory tests used to

define the patients, or

Was the assign
randomized? Randc
cent chance of bei
armed study (one
ted to the study befc
How was the randc

If the subjects t
"If it is a nonran
some extraordina

Was there a c
always be weake
son? In other wc
what are you comp
control group simil
studied?

Vogt calls "
bly . . .
ular litera

Do you have
ative of the generc
the disease or con
long way towa
are the results ap

If your grou
important popula
statistical adjustr
specific groups, o
ple, to make a
nearly compar
bility and stra

Was the st
treatment with c

25030125 19

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

define the patients, or were clinical diagnoses (necessarily less reliable) used?

Was the assignment of subjects to treatment or other intervention randomized? Randomization should give every patient a 50 percent chance of being assigned to one group or the other of a two-armed study (one comparing two groups). *Were the patients admitted to the study before the randomization?* This helps eliminate bias. *How was the randomization done?*

If the subjects weren't randomized, why not? One statistician says, "If it is a nonrandomized study, a biased investigator can get some extraordinary results by carefully picking his subjects."

Was there a control or comparison group? If not, the study will always be weaker. *Who or what were your controls or bases for comparison?* In other words: *When you say you have such and such a result, what are you comparing it with? Are the study or patient group and the control group similar in all respects but the treatment or other variable being studied?*

Vogt calls "comparison of non-comparable groups probably . . . the single most common error in the medical and popular literature on health and disease."[2]

Do you have reason to believe your subjects and controls were representative of the general population? Or the particular population—those with the disease or condition you are interested in? The answers here go a long way toward answering these questions: *To what populations are the results applicable? Would the association hold for other groups?*

If your groups are not comparable to the general population or some important populations, have you taken steps to adjust for this? Either statistical adjustment or stratification of your sample to find out about specific groups, or both? Samples can be adjusted for age, for example, to make an older- or younger-than-average sample more nearly comparable to the general populace. (More on applicability and stratification after a bit.)

Was the study blind? In a study comparing drugs or other forms of treatment with a placebo or dummy treatment, did (1) those administering

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

2503012520

the treatment, (2) those getting it, and (3) those assessing the outcome know who was getting what, or were they indeed blinded, knowing only that they were comparing A and B (or A, B, and C, perhaps)?

Could those giving or getting the treatment have easily guessed which was which by a difference in reaction or taste or other results?

Not every study can be a blind study. One researcher says, "There can be ethical problems in not telling patients what drug they're taking and the possible side effects. People are not guinea pigs." True enough, but a blinded study will always carry more conviction.

Were there other accepted quality controls? For example, making sure (perhaps by counting pills or studying urine samples) that the patients supposed to take a pill really took it.

Were you able to follow your protocol or study plan?

If there were questionnaires, interviews, or a survey: *Were the questions likely to elicit accurate, reliable answers? Was it really possible to get accurate answers to these questions?*

Sampling is as common in medical studies as in political polling. Every study examines a sample, not the whole population. The sample must be reasonably accurate to give valid results. But badly worded questions can also distort the results. Respondents' answers can differ sharply, depending on how questions are asked. Example: In one study 1,153 subjects were asked which is safer, a treatment that kills 10 percent of every 100 patients or a treatment with a 90 percent survival rate? More people voted for the second way of saying precisely the same thing.[3]

People commonly give inaccurate answers to sensitive questions, such as those about sexual behavior. They are notoriously inaccurate in reporting their own medical histories, even those of recent months.

Ask: *Did you pretest your questions for effectiveness before doing your actual survey?*

Also: *What was your nonresponse rate? Do you report it?*

In any study:
course? Do you accou
Every study
David Sackett say
reasons. Rather,
recover, die, or re
ability." If an inve
dropped out, it ca
died of "other cau
being investigate
after all, they di
treatment look b
deaths in every l
Sackett add
original inceptio
more are not ac
worth reading."
"Generally true,
Professor W
few relat
contain all re
sometimes been
. . including or
what extent has
data? . . . It is
data to make t

How long u
ily survive with
really know the c
And: *Hou*
biasis—a dise
made by findi
but a cure wa
"It does pay t

25030125 21

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

In any study: *How many of your study subjects completed the course? Do you account for those who dropped out and tell why they did?*

Every study has dropouts. McMaster University's Dr. David Sackett says, "Patients do not disappear . . . for trivial reasons. Rather, they leave . . . because they refuse therapy, recover, die, or retire to the Sunbelt with their permanent disability." If an investigator ignores those who didn't do well and dropped out, it can make the outcome look better. If those who died of "other causes" are listed among "survivors" of the disease being investigated—this is sometimes done on the theory that, after all, they didn't die of the target cause—it can make a treatment look better unless there are equal numbers of such deaths in every branch of the study.

Sackett adds, "The loss to follow-up of 10 per cent of the original inception cohort is cause for concern. If 20 per cent or more are not accounted for, the results . . . are probably not worth reading."[4] (On which Dr. Thomas Vogt comments, "Generally true, but utterly dependent on the situation.")

Professor Warren Burkett of the University of Texas adds a few related and pointed questions: "*Does the paper or publication contain all results of all experiments?* Support for a hypothesis has sometimes been made to seem stronger by selective reporting . . . including only the data that most closely fit the theory. *To what extent has the data offered . . . been smoothed from the raw data?* . . . It is not unknown for researchers to clip and round data to make them fit [their] predicted results" (italics mine).[5]

How long was the study's follow-up? How long do patients ordinarily survive with this disease? Were your patients followed long enough to really know the outcomes, good or bad?

And: *How thorough was the follow-up?* In one report on amebiasis—a disease caused by an amoeba—the diagnosis was made by finding the amoeba in one of three consecutive stools, but a cure was declared after observing just one negative stool. "It does pay to read with care," a medical professor observes.

2503012522

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

Could your results have occurred just by chance? Have any statistical tests been applied to test this?

Did you calculate a P *value? Was it favorable—.05 or less?* (Reported as <.05; see Chapter 3.) *P* values and confidence statements need not be regarded as straitjackets, but like jury verdicts, they indicate reasonable doubt or reasonable certainty.

Remember that positive findings are more likely to be reported and published than negative findings. Remember that a favorable-sounding *P* value of <.05 means only that there is just 1 chance in 20, or a 5 percent probability, that the statistics could have come out this way by pure chance *when there was actually no effect*—so 1 in every 20 statistically significant results may be a misleading false positive.

There are also ways and ways of arriving at *P* values. For example, an investigator may choose to report one of several *end points:* death, length of survival, blood pressure, other measurements, or just the patient's condition on leaving the hospital. All can be important, but a *P* value can be misleading if the wrong one is picked or emphasized.

You might want to ask: *Are all the important end points and their* P *values reported?* Also: *Was the test giving the* P *value the appropriate test, as planned in your written protocol, or did you finally do more than one kind of test?* (And perhaps report only the best answer?) *What were the other values?*

Did you collaborate with a statistician in both your design and your analysis? A statistician's collaboration often may be indicated in a credit or footnote.

In studies seeking *cause and effect,* remember that association is not necessarily causation. Rutgers' Dr. Michael Greenberg reminds us, "Mathematical methods cannot establish proof of cause and effect. They can indicate the probability that a relationship occurred by chance, can sometimes quantify the existing relationship between actions and effects, and can under the best circumstances be used to predict the impact of actions even

25030125 23

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

if the complex phenomena driving them are not understood. . . . View mathematical associations with a healthy degree of skepticism."

A true experiment, controlling all variables, can sometimes prove cause and effect almost surely. This is easier in physics and chemistry than in human biology. When, then, does a close association in an observational study (rather than a controlled experiment) indicate causation? There are several possible criteria that you can ask about:

Is the association consistent? Are similar results usually found in different places and by different research methods?

How strong is the association? If risk is an appropriate way of describing a particular situation: *What is the relative risk, or the risk ratio?* The word "strong" is used here in its mathematical sense. It mainly means the *magnitude* of an effect or risk, the *odds* favoring the *outcome of interest* versus no such outcome.

A *relative risk,* or *risk ratio,* compares two rates by dividing one by the other. In an American Cancer Society smoking study (see page 46) the lung cancer mortality rate in nonsmokers aged 55 to 69 was 19 per 100,000 per year; the risk in smokers was 188 per 100,000. Since 188 divided by 19 equals 9.89, the smokers were about 9.9 times more likely to die from lung cancer—their relative risk was 9.9.[6] That's strong!

Is there an impressive dose-response, or cause-and-effect, curve—a curve or *gradient* that shows that the greater the exposure to the agent, or cause, the greater the effect? Heavy smokers are indeed at greater risk than moderate smokers, and moderate smokers at greater risk than light smokers. (In some cases—this is an unsettled matter—there may be a threshold effect, an effect only after some minimum dose.)

Another way of asking about risk and response: *What is the correlation coefficient*—the extent to which a set of measurements of the association is *linear?* A perfect linear relationship, or correlation, between two observations or variables would show up as a straight, steadily rising set of *data points*—in everyday language, a straight line on a graph. A perfect positive correlation or,

2503012524

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

linear relationship, is given the value +1; +.5 would be a lesser but still interesting relationship; −1 or any negative figure indicates an *inverse* or *negative relationship,* such as a runner's speed going down as his weight goes up. A correlation of zero means no consistent association.

How specific is the association? Does a supposed cause lead to many supposed effects? Or does an effect depend on many supposed causes? Such associations are less specific, and thus more suspect, until positive evidence piles up. Smoking indeed causes many effects. A lung disease, asbestosis, is most common when there is exposure to both asbestos and cigarette smoke.

Does the supposed cause precede the effect? Is a supposed biological association epidemiologically plausible? One strong argument for a cause-and-effect relationship between high consumption of saturated fats and cholesterol and coronary heart disease is that populations on such diets generally develop more such disease than those on leaner diets.

Does the association make biological sense? Does it agree with current biological and physiological knowledge? You can't follow this test out the window. Much biological fact is ill understood. Also, Mosteller warns, "*Someone* nearly always will claim to see a [biological or physiological] association. But the people who know the most may not be willing to."[7]

Finally, look for the real why. Ask: *Are there other possible explanations? Did you look for other explanations—confounders, or confounding variables,* that may be producing or helping produce the association? Sometimes we read that married people live longer than singles. Does marriage really increase life span, or may medical or other problems make some people less likely to marry and also die sooner? Maybe the Dutch thought storks brought babies because better-off families had more chimneys, more storks, and more babies.

Did you take steps to control or adjust for other possible explanations? Did you do a stratified analysis—a breakdown of the data by strata like sex, race, socioeconomic status, geographical area, occupation? Men commonly have more bronchitis and cirrhosis of the

liver than w
more heart
possibly beca
analyses will
Did you
ate analysis) t
analyses can
also be misu
Some sophis
analyses did y
the more an
consider? Hou
tor tries eno
tion, he or
untrue.
In caus
reanalysis of
independen
see if
or reanaly
analysis or
among auth
reasoned a
than the an

In stud
know or deci
explicit, objec
ments or te
terviews, pl
highly subj
provement
quantify) o
Was there so
If two c

25030125 25

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

liver than women because they drink more. They also have more heart disease, possibly because they've smoked longer, possibly because some hormones protect women. Only stratified analyses will bring out such differences.

Did you do an analysis (a regression or some other form of multivariate analysis) to try to identify the important variable or variables? Such analyses can often reveal the strongest associations. They can also be misused, and they are not always needed or appropriate. Some sophisticated questions, when appropriate: *How many such analyses did you have to run to decide on the appropriate one?* Sometimes the more analyses, the worse the study. *How many variables did you consider? How many of these did you wind up reporting?* If an investigator tries enough variables in a kind of statistical fishing expedition, he or she is almost bound to find something, true or untrue.

In cause-and-effect and other studies, ask: *Has there been any reanalysis of the data?* "Results, if possible, should be method-independent," Greenberg believes. "You should recalculate and see if the results hold up."

A word of caution: Questions about multivariate analyses or reanalyses can be tricky. Whether or not to do one kind of analysis or reanalysis or none at all is often a matter of dispute among authorities. Launch the subject with some humility. A reasoned answer, affirmative or negative, may tell you more than the answer's precise content.

In studies of medical treatments or preventives: *How did you know or decide when your patients were cured or improved? Were there explicit, objective outcome criteria?* That is, were there firm measurements or test results rather than physicians' observations in interviews, physical examinations, or chart reviews, all techniques highly subject to great *observer variation* and inaccuracy? If improvement or relief from pain—a particularly soft (hard to quantify) outcome measure—had to be judged by observers: *Was there some systematic way of making an assessment?*

If two or more groups were compared for survival, was their starting

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

2503012526

point the same at onset? At diagnosis? At start of treatment? Were they judged by the same disease definitions at the start and the same measures of severity and outcome?

Did the intervention have the good results that were intended? Has there been an evaluation to see whether it was a useful result?

Investigators often report that a drug or other measure has lowered blood cholesterol levels. Fine, but were they able to show that it reduced the number of heart attacks? Or was reduction of a supposed risk factor itself taken to mean the hoped-for outcome? That may often be necessary, but the issue should be discussed.

Investigators once reported that a new heart drug reduced the number of recurrent myocardial infarctions (heart attacks), fatal and nonfatal. But total mortality for all causes was higher in the treated group than in a placebo group.

Public health officials may announce the success of a campaign to take high blood pressure measurements: X number of people were found to be hypertensive and were referred to their doctors. But how many went to their doctors? How many of those received optimum treatment? Were their blood pressures reduced? (If they were, the evidence is strong that they should suffer fewer strokes.)

In short: *What was the bottom line? Did you really do any good?*

To whom do your results apply? Can they be generalized to a larger population? Are your patients like the average doctor's patients? Is there any basis in these findings for any patient to ask his or her doctor for a change in treatment? Clinic populations, hospital populations, and the "worst cases" are not necessarily typical of patients in general, and improper generalization is unfortunately common in the medical literature.

Again and again, in many of the cases cited in this chapter, ask: *Do other studies back you up? Are your results consistent with other clinical and experimental findings? Have your results been repeated or*

25030125 27

Source: https://www.industrydocuments.ucsf.edu/docs/gnbj0000

confirmed or supported by other studies? Or have only you been able to get these results?

Virtually no single study proves anything. Two or 4 or 15 studies add credence, especially if the diagnostic and outcome criteria and the people studied are similar. Consistency of results in humans, animals, and laboratory tests also adds credence.

One scientist warns, however, "You have to be wary about a grab bag of studies with different populations and different circumstances." To which Harvard's Mosteller adds, "Yes, be wary, but consistency across such differences cheers me up." And Dr. John Bailar tells us that, despite possible pitfalls, "*meta-analysis* of several low power reports"—that is, statistically analyzing and integrating their results—"*may* come to stronger conclusions than any one of them alone" (italics mine).[8]

Mostly just good-sense questions? Of course. Some of the most important questions of all for a reporter to ponder are these: *What do I think? Do the conclusions make sense to me?* Do the data really justify the conclusions? If this person has extrapolated beyond the evidence, has he or she explained why and made sense?*

Does the investigator frankly document or discuss the possible biases and flaws in the study? A good scientific paper should do so. *Does the investigator admit that the conclusion may be tentative or equivocal?* Dr. Robert Boruch of Northwestern University says, "It requires audacity and some courage to say, 'I don't know.'"[9] *Do the authors use qualifying phrases?* If such phrases are important, we are bound to include them in any responsible story.

Ask the investigators themselves: *How much weight should your work be given? Is it really firm? And how important?* An experienced science reporter says, "I have found that good researchers generally have an honest and proportionate view of their

*Frederick Mosteller disagrees with my occasional reference to good sense or common sense. If something is a commonsense idea, he says, "surely all would have thought of it. So it must be uncommon sense after all." He makes good sense.

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

2503012528

own work's importance." But there are many exceptions.

Ask others in the same field: *How do other informed people regard this report—and these investigators? Are they speaking in their own area of expertise, or have they shown real mastery if they have ventured outside it? Have their past results generally held up? And what are some good questions I can ask them?* True, a lot of brilliant and original work has been pooh-poohed for a time by others. Still, scientists survive only by eventually convincing their colleagues.

More formally: *Has there been a review of the data and conclusions by any disinterested parties?* Some major clinical studies are reviewed by independent second parties or committees. Reports of the National Academy of Sciences must pass muster by a review committee.

Has there been peer review of the material? That is, has it been examined by referees who were sent the article by a journal editor?

And, a very important question: *Has the work been published or accepted by a reputable journal? If not, why not?* The *New England Journal of Medicine* prints only 15 percent of the papers submitted to it (many, of course, are rejected because they are not of enough interest to the journal's readers). Many have been given at medical or scientific meetings, yet do not pass peer reviewers' or the editors' muster. Most are eventually published elsewhere, many in good journals. But there are journals and journals.

In science as a whole, including biology and often basic medical sciences, *Science* and the British *Nature* are indispensable. In general medicine and clinical science at the physician's level, the best, most useful journals are probably *New England Journal of Medicine, Journal of the American Medical Association, Annals of Internal Medicine, Canadian Medical Journal, Journal of Clinical Investigation,* and the British *Lancet* and *British Medical Journal.* There are many equally good specialty journals as well as mediocre ones. In epidemiology, three good sources are *American Journal of Epidemiology, Journal of Chronic Diseases,* and *Preventive Medicine.* Ask people in any field: What are the most reliable journals, those where you would want your work published?

25030125 29

Source: https://www.industrydocuments.ucsf.edu/docs/qnb10000

Some of the most valuable journals to a medical reporter are not journals of original publication but review publications like *Family Practice* and *Hospital Practice,* which mainly print summary articles for practitioners. With some strong exceptions, the free-circulation—also known as controlled-circulation—journals and medical magazines, which depend wholly on advertising for revenue, are not as rigorously screened as the traditional journals. They are often on top of the news, however. All journals print clinkers sometimes. "Scientific journals are records of work, not of revealed truth," says the *New England Journal*'s Dr. Arnold Relman.[10]

Read the entire journal article yourself, if there is one. Ask the investigator for a copy or phone the journal. Or, assuming the article has already been published, look for it at a medical library, which can be found at any medical college, most good hospitals, and the headquarters of many county medical societies. Too many news releases tout articles that read far more conservatively than the PR version. Many scientists go much further in interviews or news conferences than they are willing to go in their articles. A reporter asked a scientist, "Does peer review of an article put you at ease?" He said, "It should help put you at greater ease, but nothing puts me at ease until I've read the article."

Most reporters can't be scientific referees, but *when you read an article, look for the following:*

- A credit or footnote indicating collaboration with a statistician, and a paragraph describing the method of statistical analysis and its outcomes, such as *P* value or confidence level, power to detect treatment effects, and so on. If they're in place, you can at least assume that some effort was made to apply the rigors of statistical analysis. If they're missing, should you beware? Sometimes. Sometimes the statistician is a coauthor whose specialty isn't identified. And some investigators are well versed in statistics.
- Tables and figures that tell the same story as the conclusions. Sometimes they don't. One statistician told reporters,

2503012530

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

"Don't assume that someone can interpret his own data. You may do better." And "muddle around in the footnotes and appendices," Mosteller advises. "You might find a few horrors. That's how people found out that a much publicized study of public and private schools included only about 12 private, nonparochial schools."

• Other things described in this chapter, such as the protocol and study design, the criteria for admitting and randomizing subjects, the therapy actually received (in contrast to that planned in the protocol), blinding, complications, loss to follow-up, follow-up time, and any discussion of reservations or weaknesses.

Ask, when appropriate: *Where did the money to support the study come from?* Many honest investigators are financed by companies that may profit from the outcome. So are some dishonest or self-deluding investigators. But the peddler of a biased point of view is as likely to be an antiestablishment crusader—or an academic ladder-climber—as a corporate darling. Perhaps the best question to ask yourself is, Is this investigator a scientist or a salesman? In any case, the public should know any pertinent connections.

"What proportion of papers will satisfy [all] the requirements for scientific proof and clinical applicability?" Sackett writes, "Not very many. . . . After all, there are only a handful of ways to do a study properly but a thousand ways to do it wrong."[11]

Despite impeccable design, some studies yield answers that turn out to be wrong. Some fail for lack of understanding of physiology and disease. Even the soundest studies may provoke controversy. No study settles anything for all time.

And according to Sackett, some "may meet considerable resistance when they discredit the only treatment currently available. . . . Clinicians may still elect to do something, even if it is of no demonstrable benefit. Study results may be rejected,

2503012531

Source: https://www.industrydocuments.ucsf.edu/docs/qnbl0000

regardless of their merit, if they threaten the prestige or livelihood of their audience."

Reporters need to tread a narrow path between believing everything and believing nothing. Also—we are reporters—some of the controversies make important stories.

2503012532

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000

6 Tests and Testing

Testing is often the only way to answer our questions, but it doesn't produce unassailable, universal truths that should be carved on stone tablets. Instead, testing produces statistics, which must be interpreted.

—Robert Hooke

Who knows when thou mayest be tested?

—Ronald Arthur Hopwood

DO physicians always know what they're doing when they administer tests? Stanford's Dr. Eugene Robin says many tests "have not been properly evaluated and in fact may be useless or harmful." He asks, "Is it common practice in medicine to perform careful clinical trials before introducing tests that can affect the welfare of masses of patients? Sadly, the answer is no."[1]

A good test should detect both health and disease and do so with high accuracy. The measures of the value of a *clinical test,* one used for medical diagnosis, are *sensitivity* and *specificity,* or, simply, the ability to avoid *false negatives* and *false positives. Sensitivity* is how well a test identifies a disease or condition in those who have it—how well it *avoids false negatives,* or missed cases. If 100 people with a condition are tested and 90 test positive, the test's sensitivity is 90 percent. *Specificity* is how well a test identifies those who do not have the disease or condition—how well it *rules out false positives,* or mistaken identifications. If 100 healthy people are tested and 90 test negative, the test's specificity is 90 percent.

Sensitivity, in short, tells us about *disease present. Specificity* tells us about *disease absent.* A highly unspecific test will produce many false positives; a highly insensitive test, many false nega-

25030125 33

Source: https://www.industrydocuments.ucsf.edu/docs/qnbj0000